MASTER IN COMPUTER SCIENCE PROFESSIONAL FOCUS IN SOFTWARE ENGINEERING

Stable Forward Search for Feature Selection

Michel, Gauthier

Award date:
2018

Awarding institution:
University of Namur

Université de Namur
Faculty of Computer Science
Academic Year 2017–2018

Stable Forward Search for Feature Selection

Gauthier MICHEL

Internship mentor: Benoit Frenay

Supervisor: _____ (Signed for Release Approval - Study Rules art. 40)
Benoit Frenay

Co-supervisor: Benoit Frenay

A thesis submitted in the partial fulfillment of the requirements for the degree of Master of Computer Science at the Université of Namur

Abstract: In order use machine learning, a model needs to be trained based on a dataset characterized by a feature set. This set can contain numerous and not always useful features for the model. The feature selection can help to sort features and reduce the set of features. The goal is to take the most useful of them and maximize the learning phase of the model. There are different kinds of selections based on different methods that offer various benefits like rigorous or execution speed. The number of features selected need to be smaller to reduce the computation time and the complexity without losing information. The utilization of a stopping criteria is important and can make the difference between a relevant or irrelevant selection.

Abstract: Pour utiliser l'apprentissage automatisé, un modèle doit être entrainé en se basant sur un ensemble de données caractérisées par un ensemble de caractéristiques. Cet ensemble peut contenir de nombreuses caractéristiques qui ne sont pas toujours utiles pour le modèle. La sélection de caractéristiques aide à trier et réduire ce nombre. Le but étant d'essayer de prendre les caractéristiques les plus utiles et ainsi maximiser la phase d'apprentissage du modèle. Il y a différentes sortes de sélections de caractéristiques qui se base sur différentes méthodes qui offrent des avantages tel que la rigueur ou la vitesse d'exécution. De plus le nombre sélectionné doit être le plus petit possible afin de réduire le temps de calcule et la complexité sans perdre d'informations sur les données. L'utilisation d'un critère d'arrêt est importante et peut faire la différence entre une sélection pertinente et non pertinente.

First of all, I would like to thank my Internship mentor and supervisor Benoit Frenay for the time he has dedicated to providing me with the methodological tools I need to conduct this research and this master thesis.

Contents

Chapter 1

Introduction

1.1 Context

In today's digital world, there is more and more data and we have to deal with it. Machine learning processes this huge amount of data and uses them in different ways. During the machine learning process a model is trained and from this model different tasks can be done like classification or regression on dataset. The model needs to be trained from a set of features that characterizes a dataset. These features can be information about a patient, some observations about specific element and more. The more features a dataset have, the longer the model takes to be trained. There are datasets with reasonable amount of features (less than 100), but the number of features increases constantly and modern datasets have more than 1000 features.
The number of features has some impact on machine learning process and can be a problem when there are many features. On the hand, the number of features influences the time needed to train a model and it takes more time where there are lots of features. On the other hand, a higher number of features increases the risk of noise and overfitting because there are more risk of redundant or irrelevant features. The noise and irrelevant features have negative impact on the training and the predictive power of the model.
Then it exist solutions like feature selection (FS) that helps to reduce the feature set used to train a model. It reduces the complexity of data, help to understand data and reduces the computation time needed to train the model. FS aims to take features that are the most useful for training to reduce the noise in data and reduce the risk of overfitting.
An important concern in FS is to realize a stable and reliable feature selection. The FS is stable if it chooses same features each time it is used on a specific dataset. If the FS is used on a specific dataset several

time and each subset of selected features are composed with different features it is totally useless and a waste of time. There are a number of different definitions in the machine learning literature for what it means to be a "relevant" features. This relevance of a feature depends on the target and the goal of the feature selection There are different kinds of FS that are performed at different step of the machine learning process. Wrapper methods use the model information and work directly with values processed by a model like the training accuracy. Filter methods process ahead and reduce the features set before the training of a model. They use statistical algorithms to determine which features to keep and reject. The selection can be processed by a exploring method like forward search that begin from a set and features are added one by one. Backward search begin with all features and irrelevant features are rejected. A mix of both exist and process by adding or rejecting features to find the best subset.

1.2 Master thesis goals

It is important to have a FS that is reliable and stable. The goal of this master thesis is to research on a feature selection method with a balance between this two requirements. Several research questions have been made on different aspects of the feature selection for this purpose.

The first one is to compare the performance between two kinds of selection performed in feature selection. The research question is to see how the feature selection behaves and if the selection is meaningful.
The selection method is a wrapper in forward search with two selections. One often used with wrappers that is based directly on the training accuracy of the model. The second one is based on a cross validation (CV) accuracy.

The second research question is focused on stopping criteria used to determine the best number of features to keep. The goal is to proposed new stopping criteria based on algorithms or statistical formula. These criteria is adapted to work as a threshold to stop the FS when it is not worth to add new feature to the subset of selected features. All stopping criteria are compared and ranked based on stability and relevance metrics.

Finally, the last research question is to compare a filter method with results from the wrapper. The goal is to see if results obtained for the two others questions with the wrapper are the same for the filter method.

In this purpose, the filter also uses training accuracy and CV accuracy to select features and all stopping criteria proposed.

1.3 Plan

The master thesis is separated in two section where the first is focus on the selection method and the second on stopping criteria. For the first section the chapter 2 is a state of art of feature selection to introduce technicals aspects. The chapter 3 explains metrics used to compare results of experimentation. The section finish with chapter 4 and 5 that explain the experimentation and results obtained.
The second section focus on aspect and elements used for the second research. The section begin with chapter 6 and 7 those present Hoeffding and the permutation test used to build new stopping criteria. Chapter 8 is the state of art of stopping criteria. Chapter 9 explains the adaptation of Hoeffding and permutation test to make new stopping criteria. Chapter 10 and 11 are about experimentations and results obtained for the section.
Finally the chapter 12 is a conclusion about the master thesis, discussion about research questions and future works.

Chapter 2

State of the Art

This chapter explains some important concepts that have been used. On the hand the feature selection, exploring methods used to create subsets and method used to evaluate these subsets. On the other hand, machine learning and model used for experimentations.

2.1 Feature selection

Feature selection (FS) [17] [18] consists in building a set of selected features from a set of input features of a dataset to maximize the prediction power of a model. From a set of features $\mathcal{X} = (X_1, ...X_d)$ and a target Y that has to be predicted by elements from $\mathcal{X}$. FS has to find the best subset of features from $\mathcal{X}$ that are the most relevant to predict the value of target Y. The selection of features can be processed by a wrapper or filter method.

Stopping criteria chooses the number of features in the subset of selected features. This is important because if too much features are selected by FS it is a waste of time because the purpose of FS is to reduce the features set. If you select not enough that impact the deduction power of your model. The ideal case is to stop when you reach a value close to the maximum of your heuristic (for example training accuracy).

The selection of relevant features [15] is important in feature selection, but what is a relevant feature for a feature selection ? For the model, a relevant feature provides useful information to train correctly the model while an irrelevant feature increases the noise and impoverishes the accuracy of the model. It is a central problem in machine learning, and many induction algorithms incorporate some approach to address it. Experts

need features to be relevant for the domain covered by the dataset. For example, in medical domain some features are obtained by experiences. These experiences have a price, take a certain amount of times, ... It can be more relevant to experts to have features that provide the same amount of information to the model obtained by more easily experience than by a difficult one. Only the relevance for the model is treated during the master thesis.

Stability of FS is sensitive to small perturbations in the training set. This issue is of course extremely relevant with small training samples. If changing or removing one training sample have a significant impact in feature selection this one cannot be considered as reliable and stable. So it is a important metric to evaluate to ensure the better and reliable FS.

The FS used exploring method to create and generate subsets of different size. These subsets are evaluate by wrapper or filter method to give to each subsets a score. The subset with the best score is selected as the best subset of features.

2.1.1 Exploring Method

Exploring method explains how to select or create subset of features for FS. To know which subset to keep FS needs to evaluate and scores each subset formed by exploring methods.

Forward selection [16] begin with a empty subset $\mathcal{S}$ of selected feature f. For the first step features are tested one by one with the model. The best feature f_i is selected and added to the subset of selected features $\mathcal{S}$. Then each subset of size two are formed with the previously selected feature and another feature remaining. The best pair that give the best performance is selected and saved as the best subset of size two. This process is repeated until the FS considers there are enough features for the subset of selected features.

Backward elimination Begin with the full set of input features. At each iteration the less significant feature is removed from the set. This process is repeated until there is no improvement for the model in removing another feature.

Recursive Feature elimination is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at

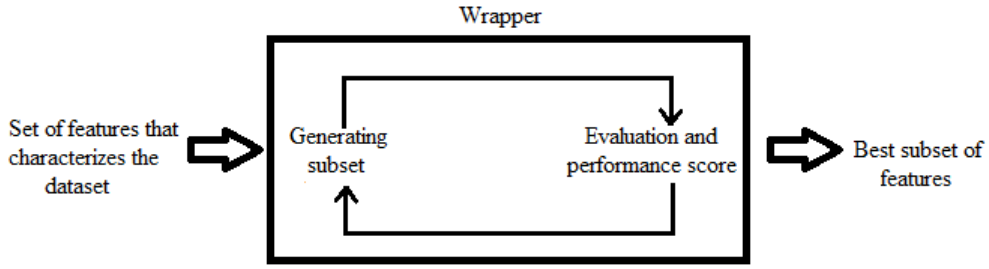


Figure 2.1: Wrapper process

each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

2.1.2 Wrapper Method

Wrapper [1] [2] [6] is a method that searches for the best subset of features by evaluating different subset of features to train the model. Wrapper uses induction algorithm from the model to determine features to select. The principle is to test different size of subset generated by a exploring method with different combination of features and training the model with each subsets. From training, the wrapper attributes a score to each subset. From scores a subset is selected and the wrapper decide if it is relevant to add or delete feature from the selected subset.

As show on figure 2.1 the wrapper begin with the set of features that characterized the dataset. From the set, the chosen exploring method generate subset of features. The subset is evaluate from the training model and a performance score is given. The process of generating subset and evaluation is looped several times. At the end, the best subset is selected by the wrapper based on the performance score.

By testing up to d^2 possibility of subset, wrapper ensures to give the best or a subset closed to the better performance for training the model. The counterpart is that wrapper is computationally very expensive. It is interesting for little data set with small set of features but not recommended for large dataset with lots of features and samples. The more combination are possible the more wrapper takes time to evaluate and selected the best subset.

2.2 Filter methods

The filter [3][4][5][6] method is a preprocessing step performed before the learning phase of the model. The Filter method does not dependent on any machine learning model because it does not use any induction algorithm of the model. Basically, filter method takes the set of features, a exploring method generates subsets, evaluates these subsets and takes subset with the best score. The evaluation relies on a score based on general characteristics of features like distances between classes or statistical dependencies. After that the new set of features chosen by the filter is used to train the model.
Filter methods are faster than wrapper and work well on large data sets. However filter does not take into account the learning phase of the model and it is difficult to know if the subset of selected features is the best that can be obtained. Several filter methods are already used in literature :

Delta test : Delta test algorithm uses the neighbors method to evaluate the quality of a feature. The nearest neighbor (NN) of a point is defined as the (unique) point which minimizes a distance metric to that point defined like this :

$$N(i) = \min ||x_i - x_j||^2. \tag{2.1}$$

An example of distance metric is Euclidian distance. The formula of the delta test is written as follows :

$$\vartheta(X) = \frac{1}{2M} \sum_{i=1}^{M} (y_i - y_N(i))^2. \tag{2.2}$$

The delta test is used to score subsets of features generated by a exploring method. The subset with the smallest score is the best one and is used to train the model.

Correlation-based Feature Selection : Correlation based feature selection (CFS) is a filter algorithm that ranks feature subsets according to a correlation based on heuristic evaluation function. The bias of the evaluation function is toward subset $\mathcal{X}$ that contain features $x_1, x_2, ...x_n$ highly correlated with the target Y and uncorrelated with each others. Irrelevant features should be ignored because they have low correlation with the target. The algorithm can deal with redundant features that will be highly correlated with one or more of the remaining features. The acceptance of a feature depends on how well the feature predicts classes in areas of instance that are not already predicted by others features. The evaluation function is :

$$M(X_i) = \frac{kr_{cf}}{\sqrt{k + K(k-1)kr_{ff}}} \tag{2.3}$$

where M(S) is the heuristic "merit" of a feature subset $\mathcal{S}$ containing k features, r_{cf} is the mean feature-class correlation ($f \in S$) and r_{ff} is the average feature inter correlation. The numerator provided an indication of how predictable of the class is a set of features and the denominator, how much redundancy there is among features.

The fast correlated-based filter (FCBF) : FCBF is based on symmetrical uncertainty. S_U is the ratio between the mutual information and the entropy of two features, $X_i \in \mathcal{X}$ and $y_i \in Y$. Where the mutual information is defined by :

$$MI(X, y) = H(X) + H(y) - H(X, y). \tag{2.4}$$

and SU algorithm like this :

$$SU(X, y) = 2\frac{MI(X, y)}{H(X) + H(y)}. \tag{2.5}$$

FCBF is known to be good to reject redundant and irrelevant feature but cannot take in consideration relation between features.

2.3 Machine learning models

Feature selection is a method to improve machine learning process and a model is necessary to perform machine learning. Models have been chosen because they can be used for classification problems as well as regression problems. A classification problem is defined like this : from a given set of tuples (X, y) of training sample with X a vector of d features and y a discrete class label, classification search to produce from these examples a function $f(x) = y$ that will associate for a new vector X the target y associated with high accuracy. The regression problem is to produce a function f that predicts continuous value output.

k-nearest neighbors : the *knn* algorithm [22] [23] is a non-parametric method that uses as input the closest training examples from feature space. The output depends on whether *knn* is used for classification or regression. For each unlabeled test sample, the method classifies it in comparison to its neighbors. The number of neighbors (k) taken into consideration is a meta-parameter. For example on figure 2.2, the algorithm has to classify the test sample (grey) between blue and green

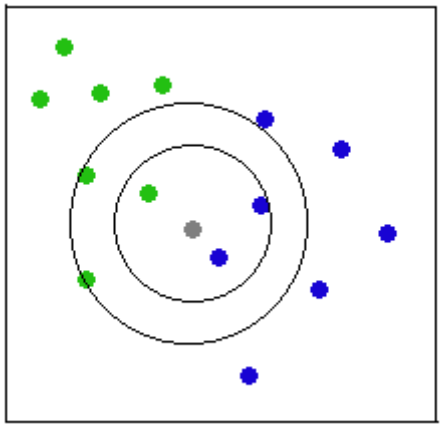


Figure 2.2: Knn example where the sample (grey) needs to be classified between blue or green elements

elements. If the number of neighbors k is 3, the sample is labeled as blue because there is more blue than green. However If $k = 5$ the sample is classified as green because there is more green than blue elements around the evaluated sample. The best number of neighbors dependent on upon the dataset but there are some heuristic techniques that can be used to choose the best number of features. The accuracy of the knn algorithm can be severely degraded by the presence of noise and irrelevant features that misdirect neighbors. If the feature scales are not consistent with their importance that impact the model accuracy. Much research has been put into selecting or scaling features to improve classification. A popular approach is the use of evolutionary algorithms to optimize feature scaling or to scale features by the mutual information of training data with the training classes. In order to determine the nearest neighbors some heuristics are used as the most of time Euclidean distance for $K - nn$.

For the experimentation knn model has been chosen. knn has as advantage to be able to handle multi-class cases, need only one meta parameter and work well with enough data. One cons is that is not easy to choose the meta parameter.

Chapter 3

Metrics

To evaluate the stability and reliability of results several metrics have been used. First the section explains metrics of reliability and then those of stability.

3.1 Reliability metrics

Mean of training or CV accuracy. Depending on the selection, there are the mean of the training or CV accuracy and its confident interval associated to it. The mean shows if the subset of selected feature proposes an interesting accuracy for the model. The best case is to reach a value close to 100 % of accuracy. It is necessary to remain cautious because values equal or too close to 100 percentile can be the result of overfitting. The consequence is that the model learns the data set instead of learning how to predict on new value. In the particulate case, the model will not be able to predict a new unknown value .

Mean of test accuracy The test accuracy is obtained by using a part of the dataset that the model does not know. Being unknown the model have to use what it learns from training accuracy to predict new value. Test accuracy is also better to see if the model has overfit. If the difference between training accuracy and test accuracy is too large (for example a training accuracy close to 100 % and test accuracy towards 60%) there is a high risk of overfitting. Test accuracy also shows how well the model works with new data and gives a more reliable idea of the true accuracy of the model.

3.2 Stability metrics

Stability metrics [9] [10] [11] [12] [13] [14] are used to evaluted if the selection is stable. The selection is stable if it gets for each repetition subsets containing the same features, selected in the same order.

Entropy : When machine learning is proceed there is something call information. This information represent the quantity of element necessary to make a prediction. For example with a tree model the information is contained in each node. A node is informative if it permits to predict samples accurately (give a accurate information). The entropy is the opposite to this information and represent how bad is the information. For a tree, the most different information gives a node, higher is the entropy and less predictive is the node.
The entropy is used as a instability metric. Entropy is based on a list that contains the percentage of selection of each feature on 100 repetitions. The entropy is proceed for each subset size :

$$H(X) = \sum_{i=1}^{n} p(x_i) \log(\frac{1}{\log p(x_i)}) \tag{3.1}$$

where X is a *vector* $\in \mathcal{X}$ and p is the probability of occurrence of X_i. Then the exponential of the entropy value is compute and report on a graphics. The less is the value of the entropy, the most stable is the feature selection. It is because if the entropy is lowest that mean that the information is high. If the information is high a accuracy decision can be done.

Consistency index [11] or **pairing** gives another useful metric of stability. The idea is to take all tuple of size k and compute the correlation between each of them to determine the average pairwise similarities between features. The formula is for two subsets A and B $\in$ X note I is :

$$I(A, B) = \frac{rn - k^2}{k(n - k)} \tag{3.2}$$

where $r = |A \bigcap B|$ is the cardinality of the intercession between A and B and k $= |A| = |B|$. Then the stability score will be computed for all subset size k for each repetition. For a given set of M sequence $\lambda = (S_1, S_2, ...S_M)$

$$\tau(\lambda(M)) = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1,\ j \neq i}^{M} I(S_i(k), S_j(k)) \tag{3.3}$$

The pairing compares all subset of same size and determines how correlated they are. If each subset has same selected features in the same order the pairing score is high. The higher is the better for the pairing metric.

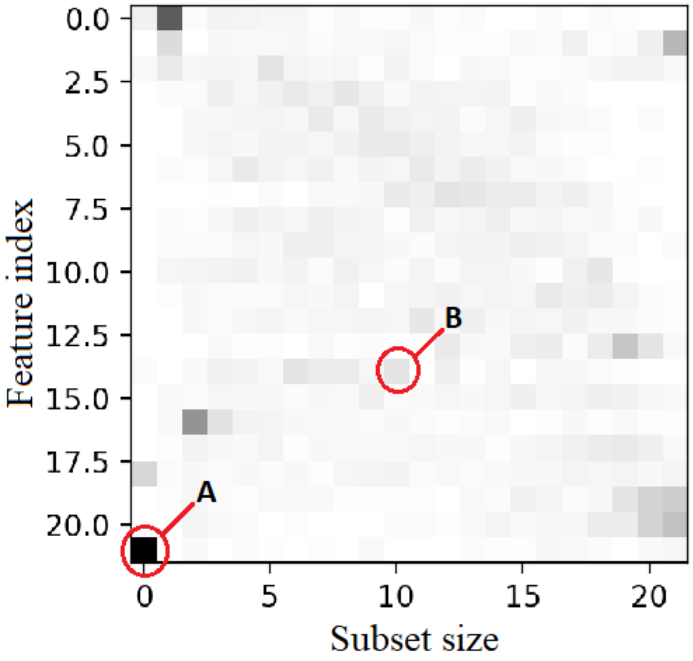
Most selected feature (max) : The last stability metric is based on color matrix. For each subset of size k, the feature the most selected is taken. It is the feature with the darker pixel on color matrix for a specific size of subset. The percentile of selection of the feature is take as a value. The higher is the value of selection the most stable is the selection because that means the same feature is selected for a subset of size k.

3.3 Additional metric

The **color matrix** is a plot (figure 3.3) that represent a matrix of pixel where each column is composed by the index of each feature. There is a column for each size of subset considered on vertical axes. Pixel intensity of a feature is directly proportional to the number of time the feature is chosen by the selection method.

$$\texttt{Darkness of a pixel} = \frac{\texttt{number of times a feature is chosen}}{\texttt{number of repetition of the experimentation}}. \tag{3.4}$$

Color matrix shows for each size of subset if the selection is stable. If a column k has one dark pixel (element A on figure 3.3) that mean that the feature links to this pixel is always chosen by features selection for subset of size k. If one feature is always chosen it means the selection for subset of size k is stable. In contrary if the column has several features colored, the selection is more unstable (element B on figure 3.3). The color matrix is also useful to highlight features that are considered as the most relevant by the feature selection. That can be convenient to analyze why a feature is chosen and if the feature selection makes sense.


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Feature index
A
B
0
5
10
15
20
Subset size

Chapter 4

Selection Comparison : Wrapper experimentation

Feature selection uses different method to reach his goal. One method is the selection of feature based on a heuristic. In the purpose of finding a reliable and stable FS it is necessary to compare different selection that can be made by a wrapper or filter. In a first time the experimentation use a wrapper to select feature and in a second time filter is used.

4.1 Recall,goal and choice of experimentation

Classic wrapper usually uses the training accuracy to determine the best subset. As training accuracy is known to easily overfit, use it to base the selection is questionable. The cross validation is known to be good to train a model. What will be the result of a feature selection that use CV score instead of training accuracy ?

Wrapper is used because it directly use learning algorithm values like training or CV score and perform an exhaustive research. The model used to perform the learning is Knn that can perform classification and regression task.

The model has only one meta parameter to manage which simplifies optimization of the training phase. Actually `gridsearchCV` [1] was used during the training task to optimize the knn and tree meta-parameter. The number of neighbors in knn help to choose at each iteration the number of neighbors to consider.

[1] a method from Scikit learn library : $http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html$

Classification datasets

Subject	samples	features
Breast cancer	569	32
Ionosphere	351	34
Sonar	208	60
wine	178	13
Parkinson	197	23

Regression datasets

Subject	samples	features
Diabetes	442	10
Boston	506	13

Figure 4.1: Classification Datasets used for experimentations found on UCI Machine Learning Repository

The experimentation was led on different classification datasets found on UCI Machine Learning Repository[2] and described in figure 4.1. Each dataset was split on two subsets, the training set with 70 percentile of the original set and the testing set with the other 30 percentile. The number of repetition for the experimentation is set to 100 to ensure the statistical soundness of experimentation.

Forward selection starts from an empty set and adds feature one by one. At each iteration a new unselected feature is added to the previous subset of selected feature of size k and a score is generated from the learning algorithm k and retain the best one.

During the first part of the experiment, the choice of the best subset at each iteration is based on training accuracy.The second part uses the cross-validation score to choose the most optimal subset of features. Cross-validation is used to perform the selection with a more rigorous parameter and the number of cross validation is set to 10.

[2]https://archive.ics.uci.edu/ml/datasets.html

4.2 Description of the experimentation

This section described the experimentation and the different steps to carry out the experimentation (Algorithm 1).
First the experimentation need to set up `Gridsearch` and split dataset into training data and test data. After the program loops d times, d being the number of features of the dataset. The purpose is to test each possible subset size. For each loop, another loop begin z time, z being the number of features not already selected by the wrapper. In this loop, training and test data subset are reduced to be as the same dimension as tested subset size. Then the model is trained with `Gridsearch` set with given parameters and the best meta parameter found. After that, the score of the subset is calculated and stored. According to the type of selection the score will be the training accuracy or the CV score. For each size of subset, the best subset is chosen which determines the feature (that is not already chosen) to add to the subset of selected features. This process is made until each subset's size have been tested. Every score of accuracy and others metrics are saved to permit the comparison between training accuracy and CV.

Algorithm 1 wrapper code

Require: $Gridsearch, Trainingsubset, testsubset$
  for each size of subset(0 to d size) **do**
    for For each remaining feature f_i **do**
      $f_i \cup$ subset of selected features
      Train model
      Compute the score of the current subset
    end for
    Select the feature f_b with the best score
    Subset of selected feature = $f_b \cup$ previous subset of Selected feature
  end for

4.3 Results

The different metrics described in the chapter 3 have been reported on various graphics.

Figures 4.2 represents the mean of the training accuracy or the cross validation accuracy score depending on what the selection is based on respectively. The mean and confident intervals are scored for each size of

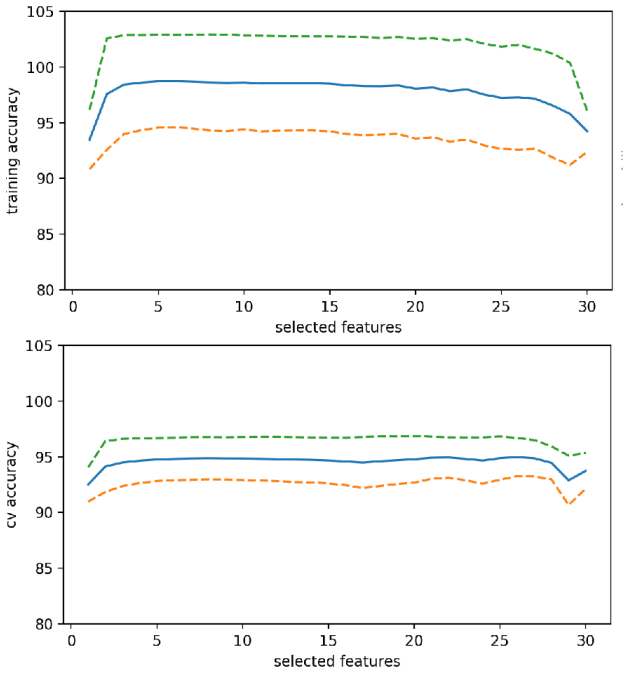
subset of selected features. With the training accuracy the stage of 100 percentile of accuracy is generally reach very fast and stay at the top. It only decreases towards the end when there are only few features left. Only the breast cancer dataset does not reach 100% of accuracy but has the same curve with a lowest accuracy. When cross validation is used it begins with less accuracy compared with training accuracy. The curve goes up to 90-95 % depending the dataset and decreases slowly after that.

Figures 4.5 represent the test accuracy for selection based on training accuracy or cross validation accuracy score. For the training accuracy, the test accuracy begin very low and rise rapidly. The increase in accuracy decreases at the end. For cross validation accuracy, the test accuracy begin higher than with training accuracy. The curve rise slowly than training accuracy to the end.

Figures 4.9 represent stability and instability of the selection. For training accuracy, metrics of stability are low at the beginning then the stability rises as the size of the subset increases. The instability is high at the beginning and decreases which is in agreement with stability behavior. For Breast cancer and Ionosphere results are different and the stability of selection stay low for each metrics. In cross validation stability metrics begin with high stability score and then decreases to be very low.

Finally figures 4.8 are representing color matrix. A diagonal tend to be drawn on the color matrix when the FS is based on training accuracy. In the case of the cross validation some features seem to be often chosen by the wrapper at the begin of the selection. After features are randomly selected which is in agreement with stability graphics. A boundary between stable and non-stable selection is clearly visible

Breast cancer



Ionosphere

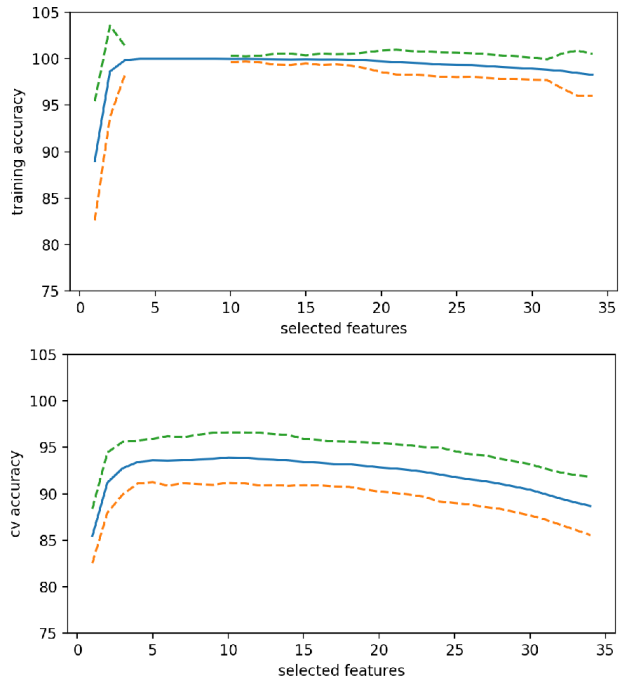


Figure 4.2: Mean of the Accuracy scores obtained with forward search where selected features is the number of features in the subset. Graphic on the top is training accuracy scores and graphic on the bottom is cv accuracy scores. Dotted line are confidence interval

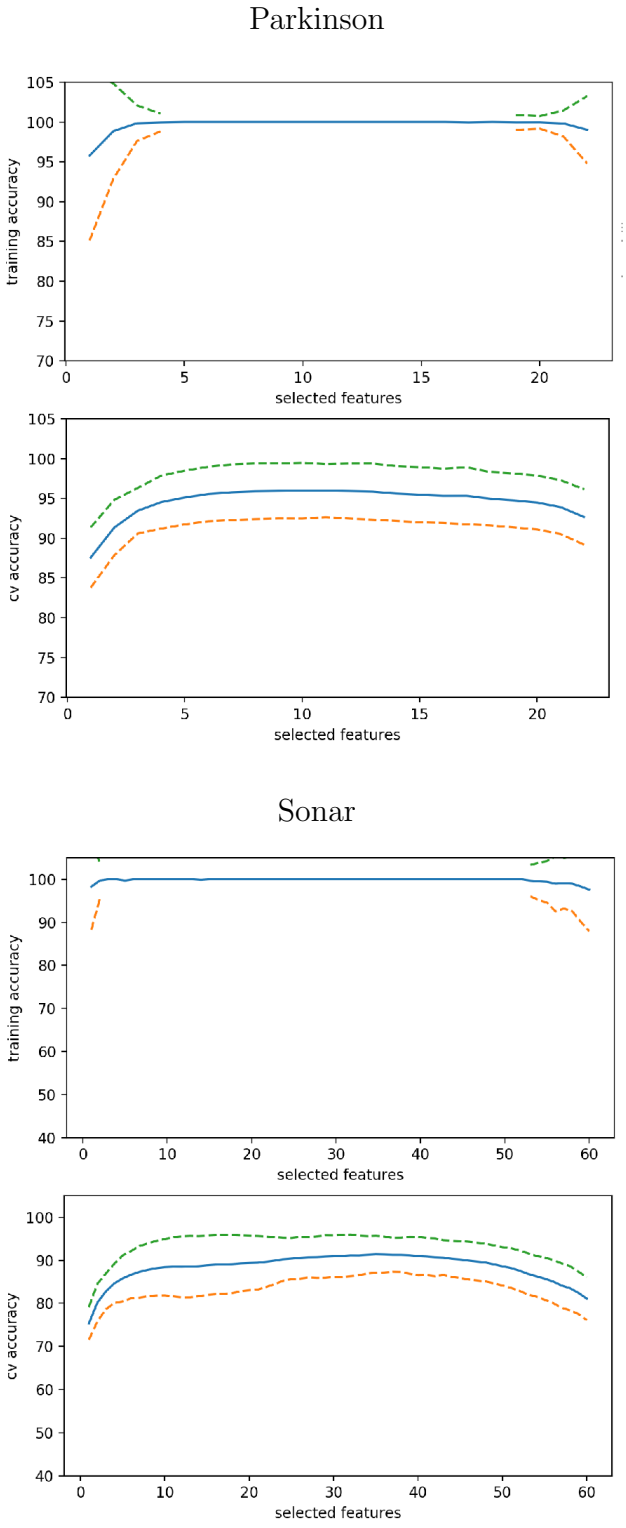


Figure 4.3: Mean of the test Accuracy scores obtained with forward search where selected features is the number of features in the subset. Graphic on the top is training accuracy scores and graphic on the bottom is cv accuracy scores. Dotted line are confidence interval.

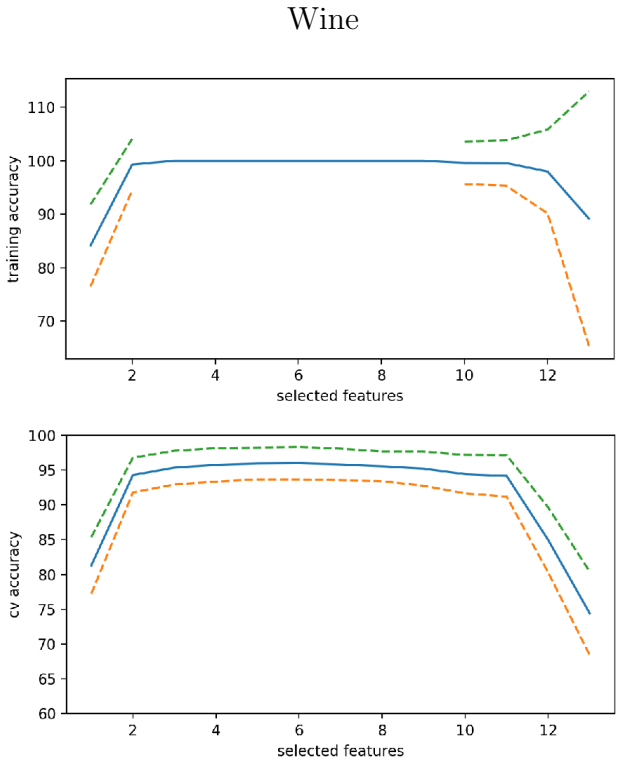


Figure 4.4: Mean of the test Accuracy scores obtained with forward search where selected features is the number of features in the subset. Graphic on the top is training accuracy scores and graphic on the bottom is cv accuracy scores. Dotted line are confidence interval.

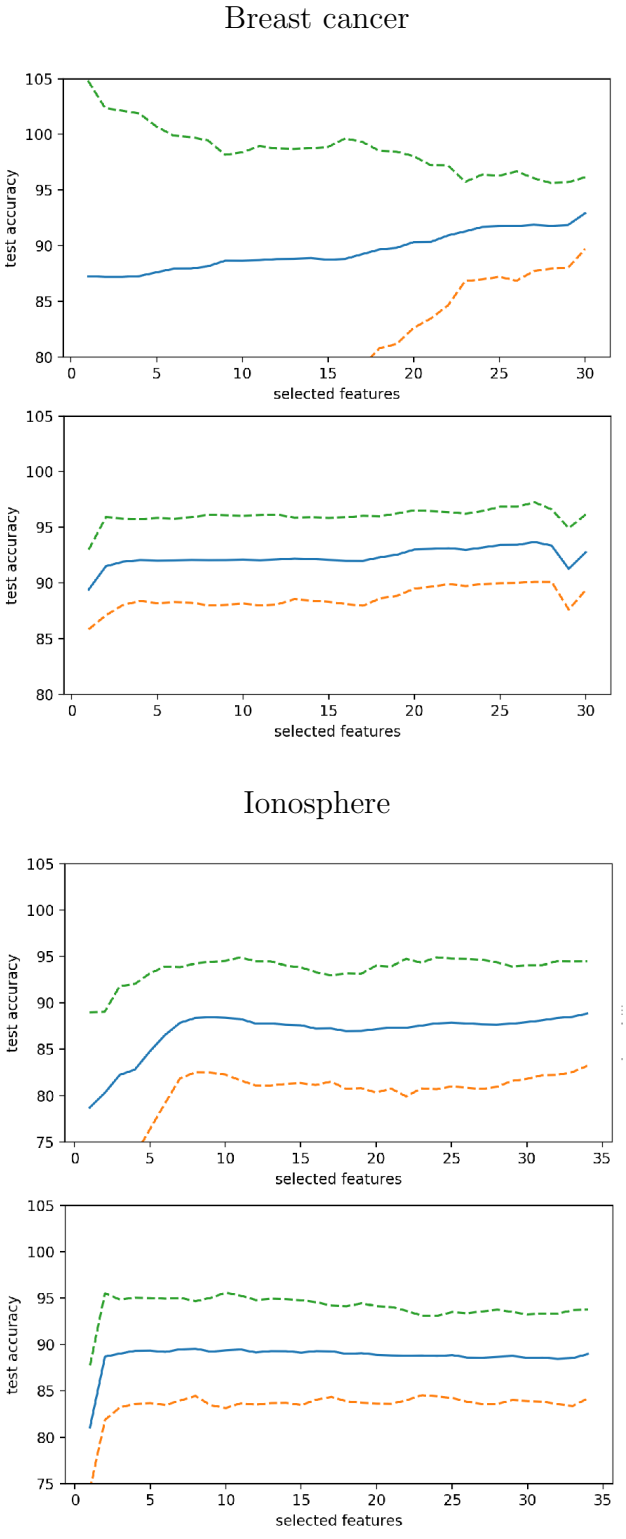


Figure 4.5: Mean of the test Accuracy scores obtained with forward search where selected features is the number of features in the subset. Graphic on the top is training accuracy scores and graphic on the bottom is cv accuracy scores. Dotted line are confidence interval.

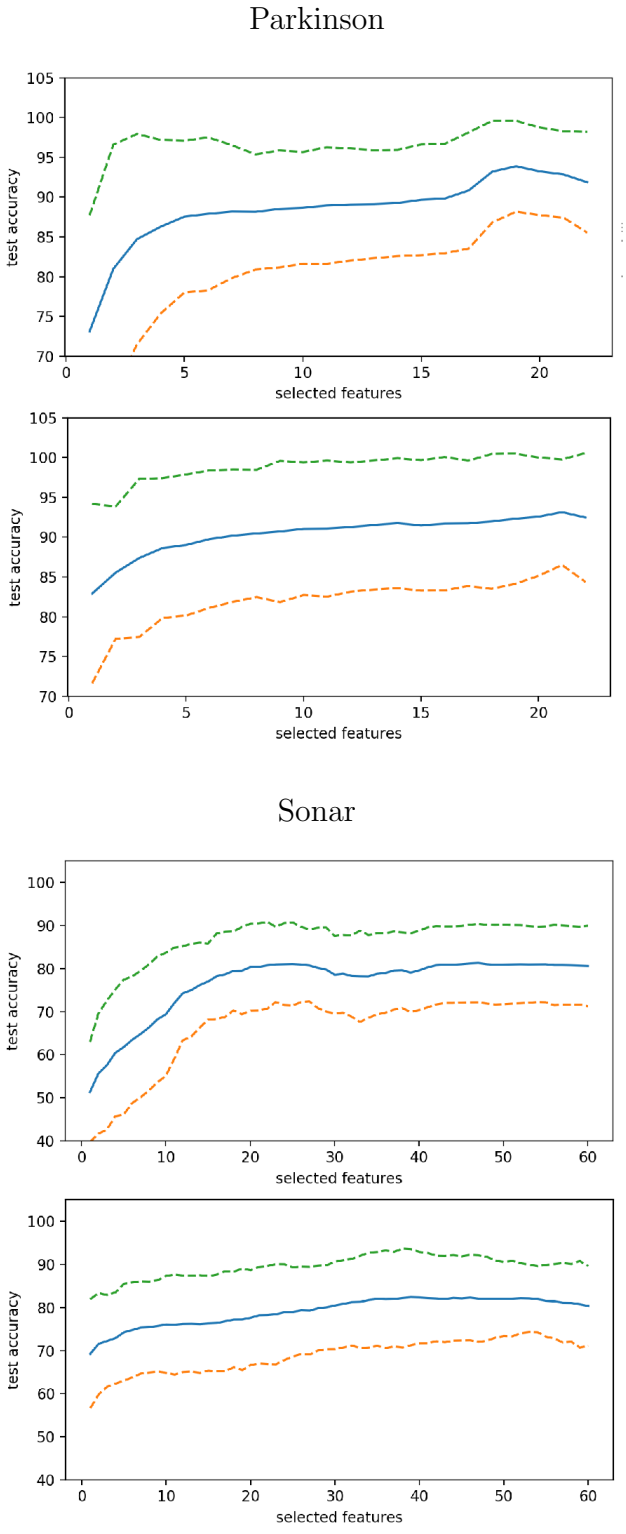


Figure 4.6: Mean of the test Accuracy scores obtained with forward search where selected features is the number of features in the subset. Graphic on the top is training accuracy scores and graphic on the bottom is cv accuracy scores. Dotted line are confidence interval

Wine

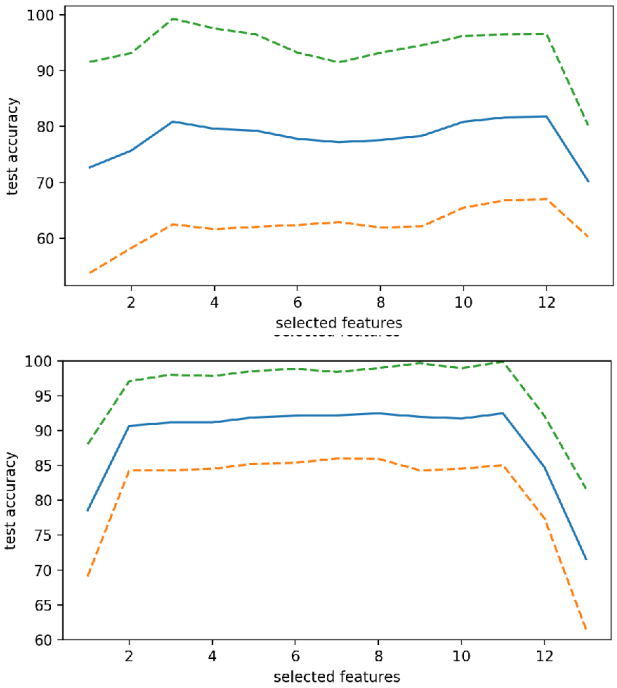


Figure 4.7: Mean of the test Accuracy scores obtained with forward search where selected features is the number of features in the subset. Graphic on the top is training accuracy scores and graphic on the bottom is cv accuracy scores. Dotted line are confidence interval

Breast cancer

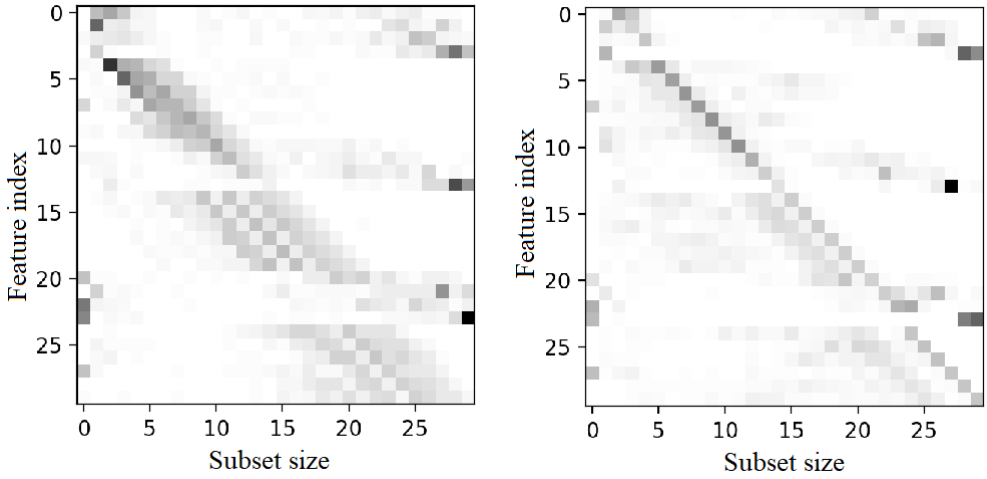


Figure 4.8: Color matrix obtained with forward search. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score

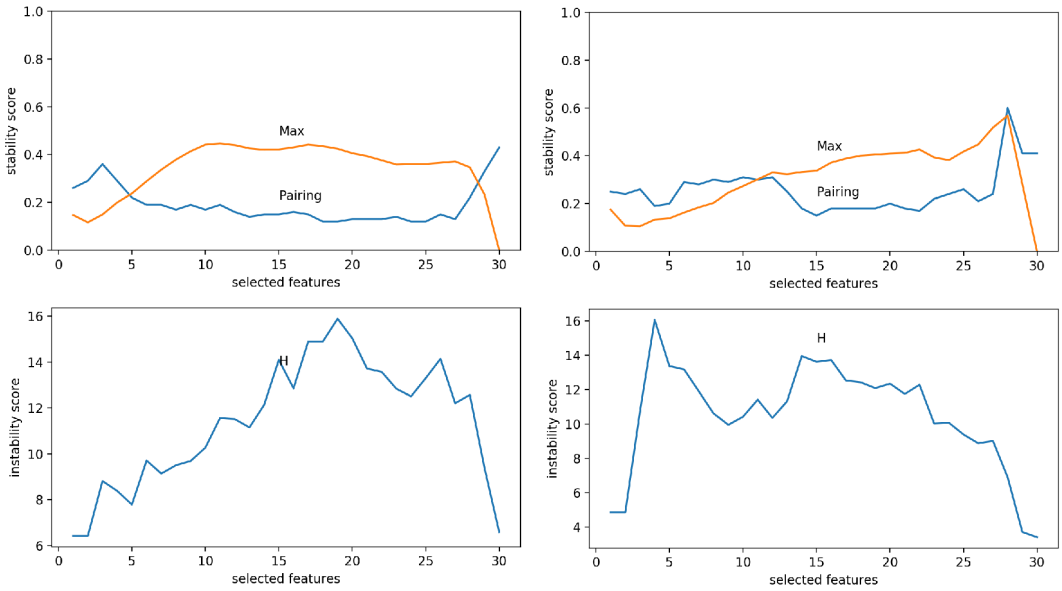


Figure 4.9: Stability scores obtained with forward search where selected features is the number of features in the subset. Graphics on the top are stability scores where the higher is the best and graphics on the bottom are instability scores where the lower is the best. Graphics on the left are experimentations with training accuracy and graphics on the right are experimentations with cross validation score

Ionosphere

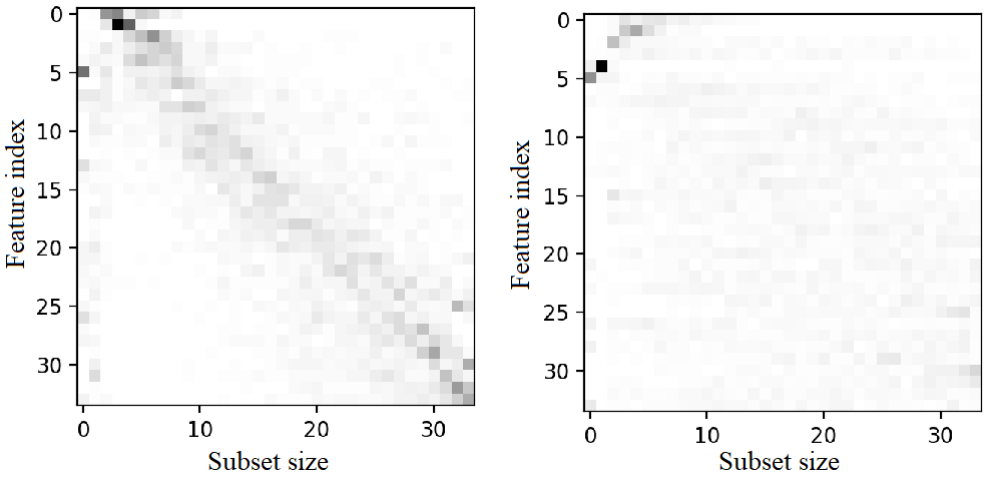


Figure 4.10: Color matrix obtained with forward search. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score

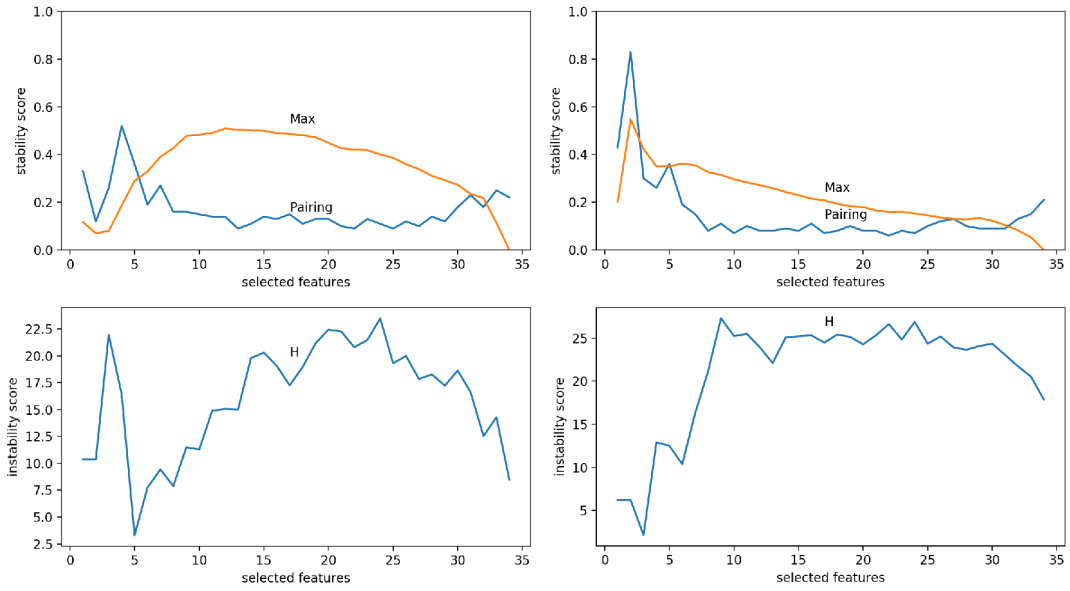


Figure 4.11: Stability scores obtained with forward search where selected features is the number of features in the subset. Graphics on the top are stability scores where the higher is the best and graphics on the bottom are instability scores where the lower is the best. Graphics on the left are experimentations with training accuracy and graphics on the right are experimentations with cross validation score

Parkinson

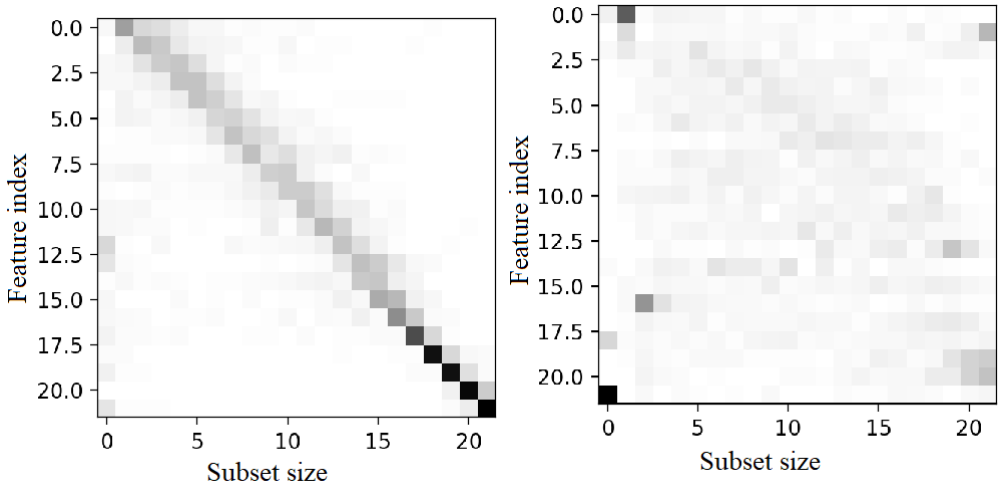


Figure 4.12: Color matrix obtained with forward search. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score

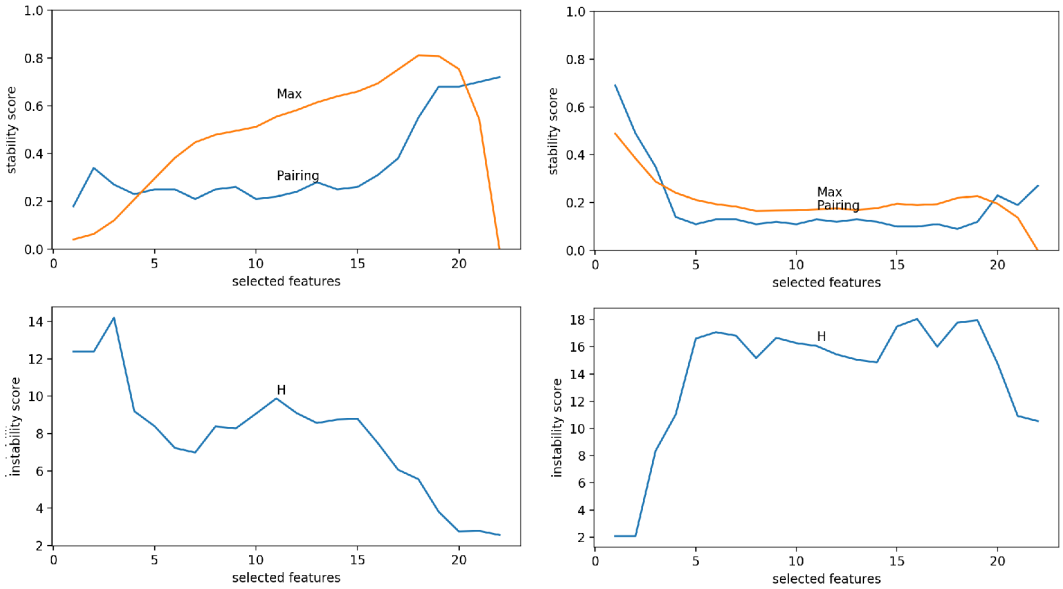


Figure 4.13: Stability scores obtained with forward search where selected features is the number of features in the subset. Graphics on the top are stability scores where the higher is the best and graphics on the bottom are instability scores where the lower is the best. Graphics on the left are experimentations with training accuracy and graphics on the right are experimentations with cross validation score

Sonar

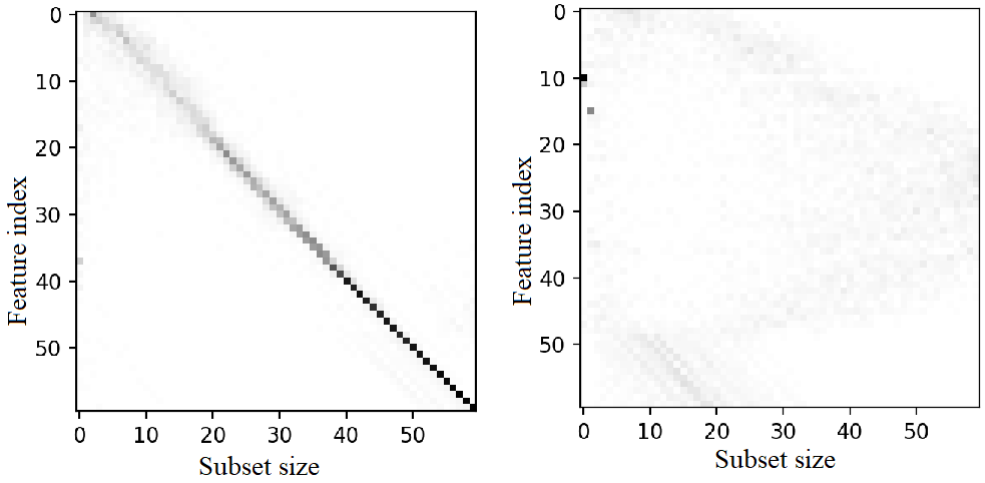


Figure 4.14: Color matrix obtain with forward search. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score

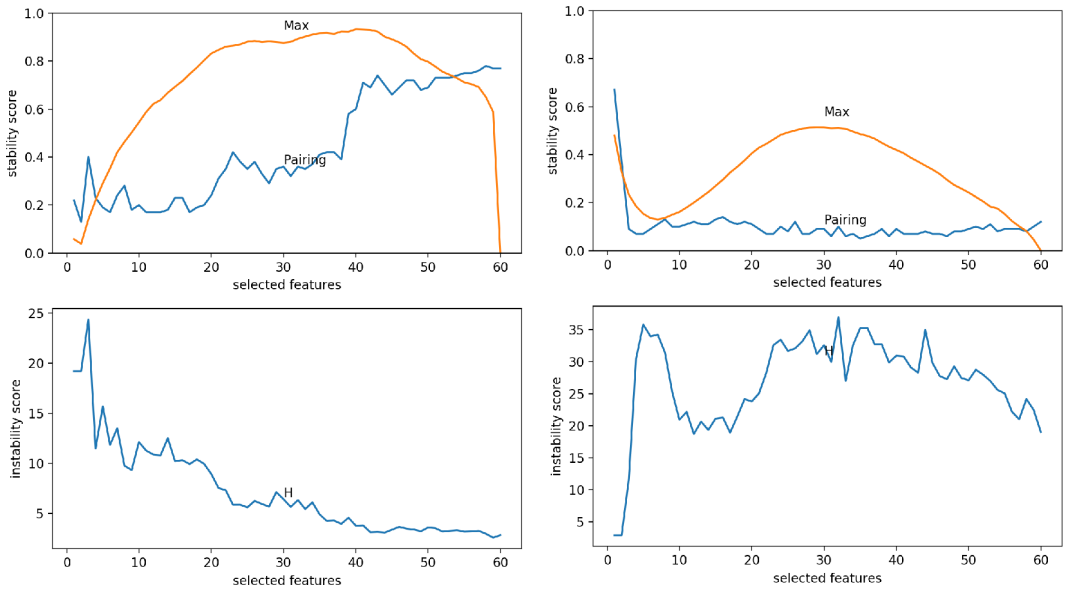


Figure 4.15: Stability scores obtained with forward search where selected features is the number of features in the subset. Graphics on the top are stability scores where the higher is the best and graphics on the bottom are instability scores where the lower is the best. Graphics on the left are experimentations with training accuracy and graphics on the right are experimentations with cross validation score

wine

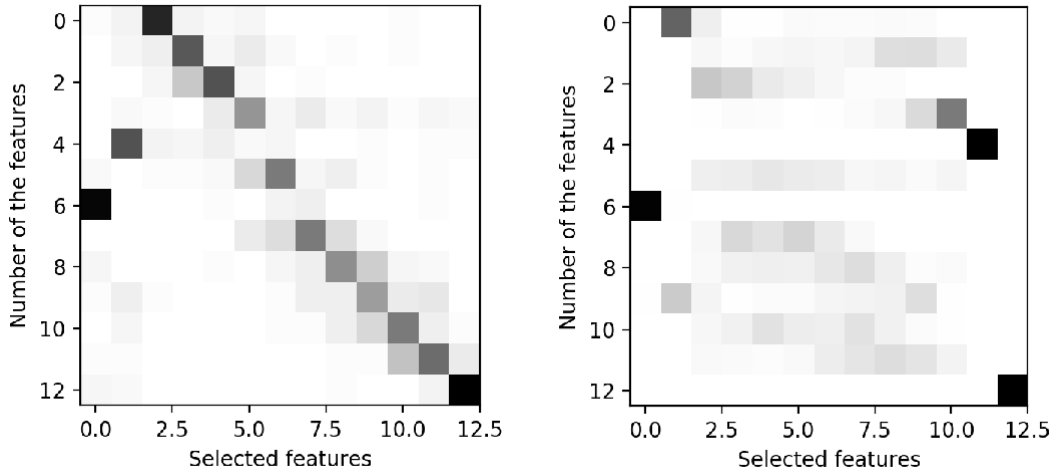


Figure 4.16: Color matrix obtained with forward search. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score

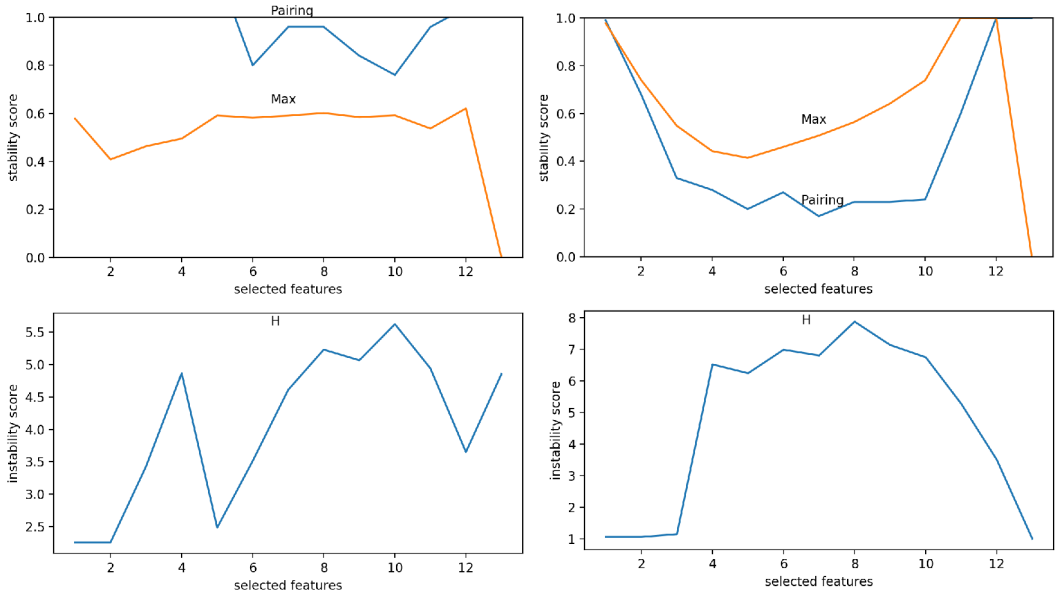


Figure 4.17: Stability scores obtained with forward search where selected features is the number of features in the subset. Graphics on the top are stability scores where the higher is the best and graphics on the bottom are instability scores where the lower is the best. Graphics on the left are experimentations with training accuracy and graphics on the right are experimentations with cross validation score

4.4 Discussion

In machine learning feature selection is an important task to improve the learning phase, reduce the cost of necessary data and the computing time needed to train model. There are several methodologies to select features like the wrapper used in the experimentation to compare the selection based on training accuracy and cross validation accuracy.

The first thing shown by results of the experimentation is the difference between results based on training accuracy and those based on cross validation shows by figure 4.3. In the first case, the training accuracy rises very fast and reach 100 percentile. It is not surprising because same data are used to train and choose features. With training accuracy, a model tends to overfit very easily and has some repercussions in the selection of features. In fact the color matrix shows on figure 4.16 a diagonal draws on it. The diagonal is obtained because the model reach 100 percentile of accuracy with the majority of features and all these features can be chosen for the next step. The wrapper takes the first feature that is not already chosen. So base the selection of the next feature on the training accuracy is inappropriate. All features reach the 100% of accuracy and for the wrapper there are all worth. The utilization of training accuracy is questionable because the training accuracy does not give the better selection of features but a default selection of them. With some datasets the wrapper seems to take a very little selection of good features before starting to select by default. It is because the cap of accuracy is not already reach and the selection was not make by default. On stability graphics 4.16 that create an artificial stability that gives the impression of a stable selection.

Selection based on cross validation seems to give more promising result with a more logical behavior based on the cross validation accuracy graphics on figure4.3. At each repetition, the wrapper often takes the same set of features at the beginning and after continues the selection in a totally random way. On the color matrix 4.16 there is generally a border between stable and unstable selection of features.The hypotheses is that the stable selection takes features considered as better or useful and the unstable selection is where the selection is totally random because not others features seems to increase the accuracy stably and get out of the heap.
Graphics based on test accuracy 4.6 give better results with CV than with training accuracy. The curve is the same but the CV accuracy has a better score of test accuracy. With a better test accuracy score, the

model has a better prediction power and can predict new feature more accurately.

Some chosen features are similar between the two criteria despite of they are less selected with the training accuracy than with the cross validation. Base the feature selection on training accuracy provide irrelevant results and the most of time can't be used in feature selection. In contrary cross validation seems to work very well and generally highlight a group of features by the frontier between stable and unstable.

Chapter 5

Selection Comparison : Experimentation with filter

5.1 Recall,goal and choice of experimentation

Filter method is another solution in feature selection that is used before the learning phase of the model. The idea is to use statistics metrics instead of induction algorithm of the model. There are classic metrics like Chi-Square, Pearson's Correlation or Delta test and more specifics metrics for feature selection like mutual information.
The goal of this section is on the hand the comparison between the wrapper method and the filter method and on other hand to compare classic Delta test with a cross validated version of it. The second goal is based on the hypotheses that using filter heuristic on all data can give distorted results. The experimentation tries to answer this question by comparing delta test with cv and without it.
For the experimentation delta test is used to compare training accuracy and cross validation accuracy. The cross validation as been set to 10 iteration and the distance between features is computed with Eucledi-ane distance. Datasets used for the experimentation as been changed from classification to regression task. Consequently the score computed by `Gridsearch` as been changed to fit with regression task. The metric choose to evaluate the accuracy of the model is $R2$.

5.2 Description of the experimentation

This subsection described how is conducted the experimentation. The first step is to prepare training and test data subset from the dataset and set all parameter that are used. Then the filter method is used to determine the best subset of features built by forward search. Then Delta test is used to evaluate each subset and gives them a score. For the cv part, the filter is used 10 times with different part of the training set and the score is computed by taking the average of the 10 scores. For each size of subset the feature or subset of features with the smaller delta test score is selected. At the first iteration, each feature is tested one by one and the best one is selected. For the others iterations, the previously selected subset combined with each remaining features is evaluated and the best subset is selected. When the subset is selected, `Gridsearch` is set up to perform to train the model with the subset chosen by the filter method.

Algorithm 2 code with delta test

Require: data subsets and parameter
  for Each subset combination **do**
    Compute delta test
  end for
  Selected the best subset x_i
  Set up Gridsearch
  Train model
  Compute scores

5.3 Results

As for wrapper, this section exposed raw results and graphics obtained with the experimentation.

Figures 5.1 show the delta test score obtain for each size of subset. With only training data for the delta test the curve rises slowly. With the CV delta test, the score begin high and decreases. For the Boston dataset the score rise a the end.

Figures 5.2 represent test accuracy score obtained by training the model with subset selected by the filer method. With only training data, the curve begin with a low accuracy. The curve increases slowly with the increase of subsets size. The curve fall at the end when it only few

features remaining. With the CV delta test, the test accuracy begin with a higher score than with only training data. Then the score increases and stabilizes when the cap of two feature is surpassed. For Boston the score decreases at the end.

Figures 5.4 represent the stability and instability of the selection. With only training data, the stability begin with a score that decreases rapidly. The max metrics is very low and fall to 0. The other maintains itself but remains low. For the entropy, the score is low but increases rapidly when the cap of two features is surpassed. With the CV delta test, the stability begin with a higher or alike score than with only training data. Then the score decreases in the majority of case but stay above the delta test without cross validation.

Figure 5.3 are representing the color matrix. Whether for the delta test with cv or only on training data the color matrix shows the same "behavior". Few features are often takes by filter at the beginning. Then the selection is more random until the end where few features are often selected last.

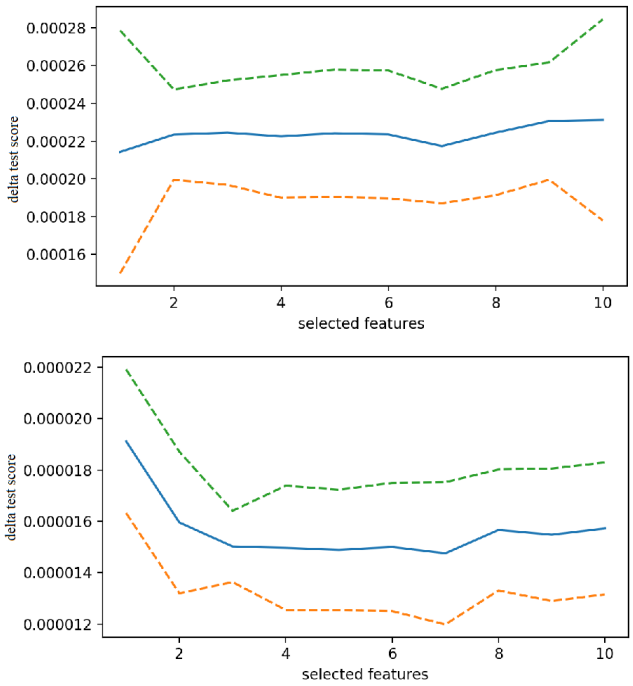
5.4 Discussion

As for wrapper the experimentation with the filter method compared scores obtained with only training data and a CV of theses scores. The purpose is to show that only using training data (training accuracy for the wrapper) gives distorted results.

First, the delta test score must decrease when the subset has more feature because there is more neighbors and the minimal distance between samples decreases. Delta score based only on training data rises with time instead of decreases show by figure 5.1. It is behavior that should not happen and that is not found with the CV version. The CV version begin with high scores then decreases to stabilize around the cap of three features. This is a more logical behavior if we refer to the delta test formula.

On test accuracy graphics 5.2 the CV delta test gives better or alike scores. That means the CV version as a model with a better prediction power. The result is the same for stability metrics 5.4 that shows stability with CV delta test aboves.

Diabetes



Boston

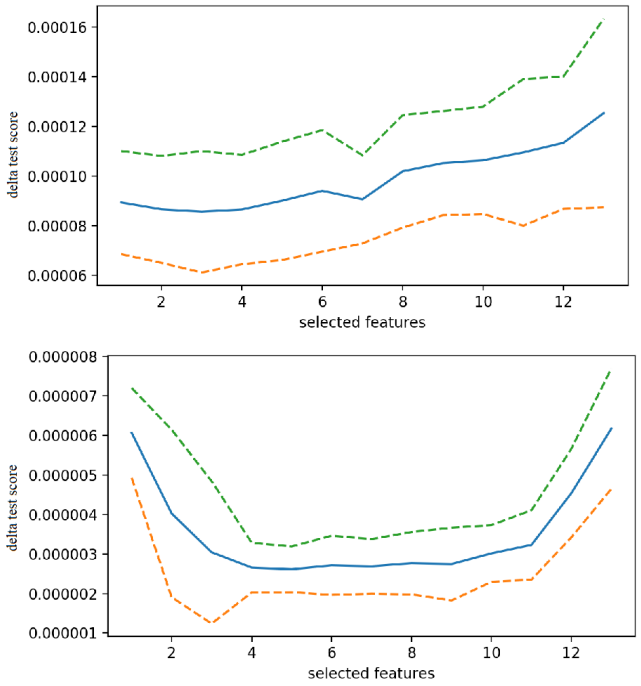


Figure 5.1: Delta test scores obtained with filter method where selected features is the number of features in the subset. Graphic on top use only training data to score delta test and graphic on bottom use cv delta test score. Dotted line are confidence interval

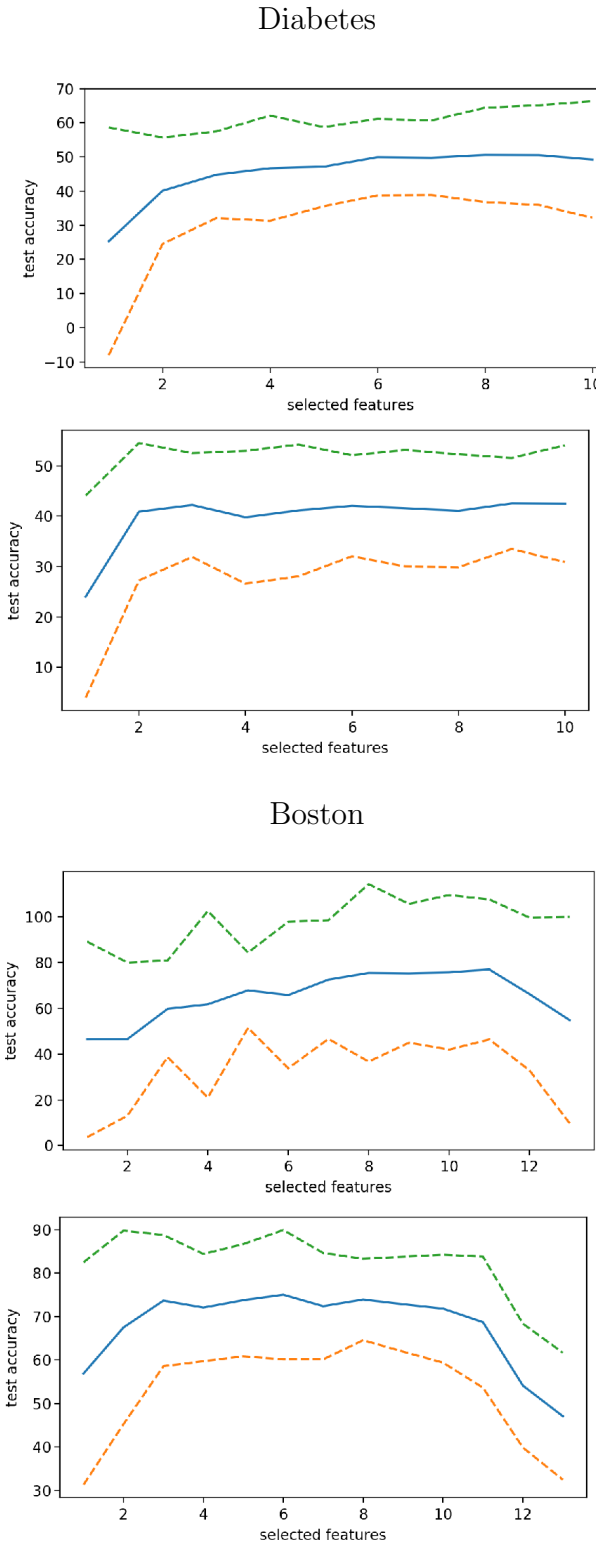


Figure 5.2: Test accuracy scores obtained with filter method where selected features is the number of features in the subset. Graphic on top use only training data to score delta test and graphic on bottom use cv delta test score. Dotted line are confidence interval

Diabetes

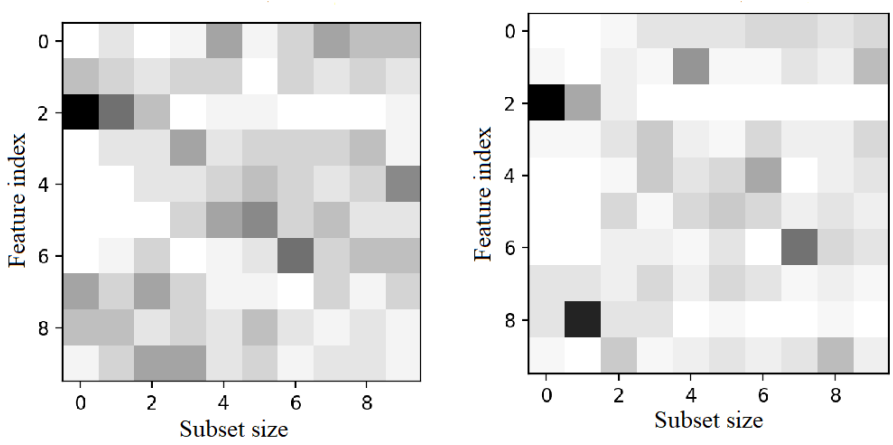


Figure 5.3: color matrix obtained with filter method. Graphic on the left use training accuracy to score delta test and graphic on bottom use cv.

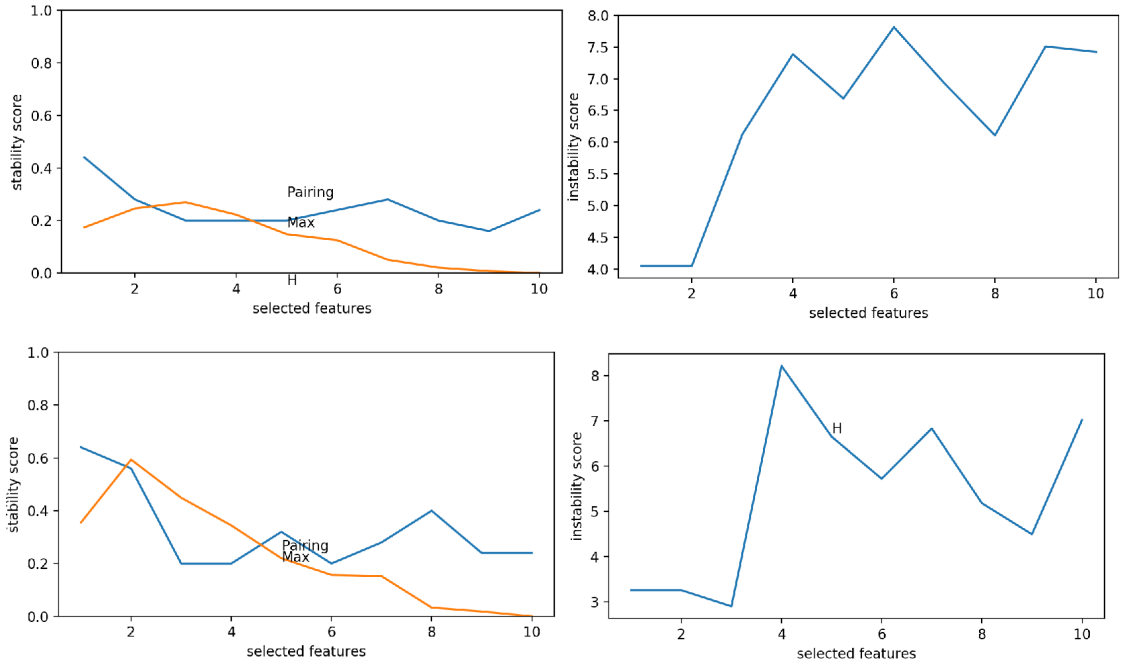


Figure 5.4: stability scores obtained with filter method where selected features is the number of features in the subset. Graphics on the left are stability and graphics on the right are instability. Graphics on the top use training accuracy to score delta test and graphics on the bottom use cv.

Boston

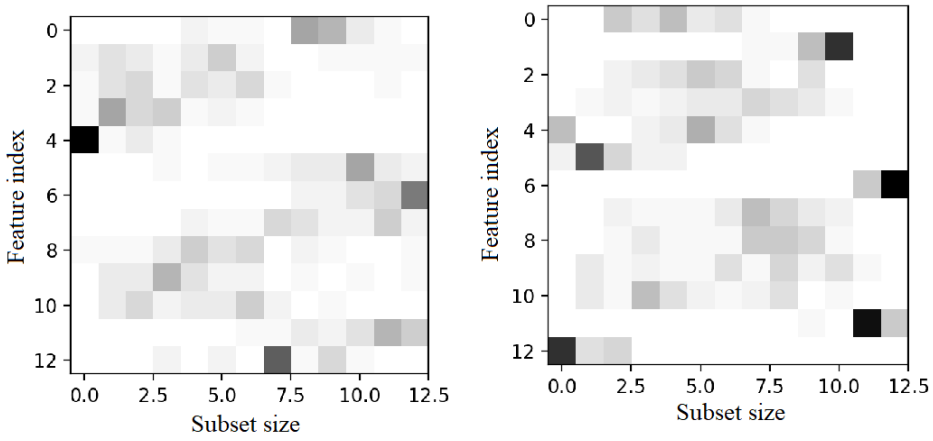


Figure 5.5: color obtained with filter method. Graphic on the left use training accuracy to score delta test and graphic on bottom use cv.

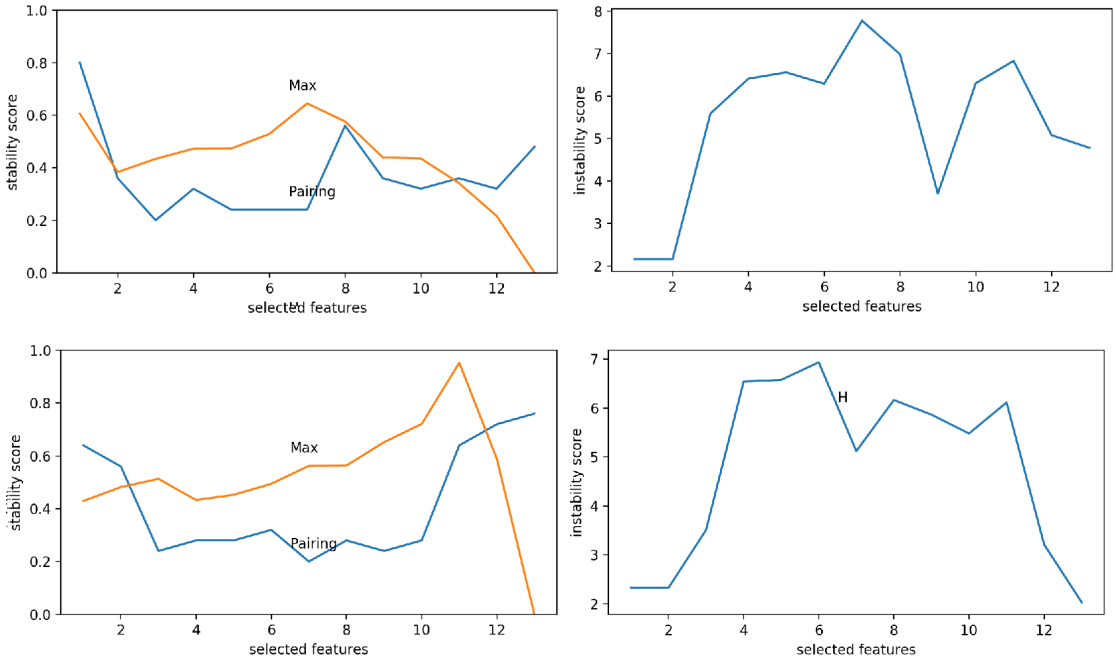


Figure 5.6: stability scores obtained with filter method where selected features is the number of features in the subset. Graphics on the left are stability and graphics on the right are instability. Graphics on the top use training accuracy to score delta test and graphics on the bottom use cv.

Color matrix 5.3 shows more similar results between the two kind of selections. They both select stably several features and after make a random selection of others features. But features selected stably are different in most case between the two methods. This shows that features selected depend on what is used to calculate the delta test and have a true impact on the selection.

As for wrapper the filter method based on CV score gives better result. It is less visible than with wrapper where the cap of 100 accuracy gives distorted results. The CV version gives results with a test accuracy alike or better that result on a more reliable model. The selection is also more stable with the CV version with metrics score aboves.

Chapter 6

Hoeffding inequality :

Hoefddings's inequality is used to provide a bound to ensure that the sum of random variable S not deviate more than real-valued of random variable. Hoeffding's inequality is usually used as a confidence interval by determining the number of samples needed to obtain this confidence interval. The inequality is :

$$P(X - E[X]) \geq e^{-2nr^2} \tag{6.1}$$

Where P is a probability, X independent random variable bounded by a interval, $0 \leq t \leq X - E[X]$ and $E[X]$ the Hoeffding's lemma.

A existing utilization of Hoeffding inequality [7] is to build a fast decision tree. In fact classic decision tree learners like ID3, C4.5 and CART assumes that all examples are stored simultaneously in memory. This constraint limit the number of training examples that can be treated at the same time. The article [7] explains that authors have designed a very fast decision tree for huge datasets. The goal being is to directly use data stream and built a potentially very complex tree with acceptable computational cost. Each node determines the number of examples needed to make the right split. When an attribute node is chosen, the succeeding examples pass down to the corresponding leaf and are used to choose to next node and so on. Hoeffding inequality is used to determine how many examples are necessary at each node to make the right decision for the split. Hoeffding ensures that with n examples, the choice made for a node is the correct choice with high probability as if the choice has been made with an infinite number of examples. So the tree is built step by step until Hoeffding considered there are not enough examples to take the correct decision for a split. To decide how many examples are necessary with a probability of $1 - \delta$ the hoeffding bound is derivative from Hoeffding inequality :

$$\epsilon = \sqrt{\frac{R^2 ln(\frac{1}{\delta})}{2n}} \tag{6.2}$$

where R is the range of a real-valued random variable X and n the number of observation of this X. ϵ is compared to a heuristic ΔG for which it is possible to make an average $\Delta\overline{G}$. If $\Delta\overline{G} > \epsilon$ Hoeffding bound ensure with a probability of $1 - \delta$ that the heuristic result is the same than a decision chosen with an infinite number of examples.

Chapter 7

Permutation test

The permutation test [8] is a non parametric hypotheses test over some estimated statistic O involving X and y. This statistic can be different according to the context for example, this statistic can be correlation, some difference between X and y, etc. If $\hat{\text{O}}$ is the value of the statistic for given X and y that are both vector of size n. The test determine how likely is the value $\hat{\text{O}}$, given the vector X and y that are supposed to be independent and thus the statistic should be zero. The empirical distribution of X, y and sample size need to be fixed. The random variable of interest is the value of the statistic O. The distribution of $\hat{\text{O}}$ is the set of all value of $\hat{\text{O}}$k for all n! possible permutation of the element of the vector x_i. The goal is to test is the proportion of $\hat{\text{O}}$k that are larger than $\hat{\text{O}}$. If more than certain % of permutation score is better than the original score the permutation test "reject" X. Make n! Permutation seem to be a little bit too much and takes lot's of computing time so an approximation with Monte-Carlo algorithm is used. The permutation can be use to set a threshold that stops the feature selection when features seem to be irrelevant. The p-value can be estimated as :

$$\alpha - 1.96 * \sqrt{\frac{\alpha(1-\alpha)}{M}} \tag{7.1}$$

There already exist several utilizations of the permutation test with others statistical metrics in order to create a threshold to stop adding feature. In this article [8], permutation test is used with mutual information. The mutual information is a statistical metric that measure how a random variable Y depend of another one X_i (Or the opposite). MI use the entropy H() and can be define like this :

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{7.2}$$

The principle is from a set X of j vector x compose with n elements. The mutual information estimator $\hat{\text{O}}$ is set. After for each vector the

estimator is computed with the x_i ,the target y and the number neighbor then M random permutations of x_i is computed. Computed estimator ô is compared with the permutation to find the percentile of permutation that are better than $\hat{\text{O}}$. If more than 5 percentile is better the feature is discard otherwise the feature is kept.

Chapter 8

Stopping Criteria

In feature selection [19] [20] [21] it is necessary to stop adding new feature to the subset of selected features when it is appropriated. In feature selection, a stopping criterion decides when to stop selecting new feature. This criterion helps the feature selection algorithm to choose a number of features that is used by the model during the learning phase. It is important in feature selection(FS) because too many features is a waste of time because FS miss is goal of reducing data features set. Not enough features impacts the prediction power of the model because there is not enough information to train the model. The contribution in this master thesis is to propose new statistical metrics that are used as a threshold of feature selection.

8.1 Classic wrapper criterion

Wrapper stopping criteria [1] [2] is also related to the model algorithm. Classic wrapper uses the value of model performance as a stopping criterion. In the case of forward selection, at each addition of a feature to the subset of selected features, the value of this subset is compared with the value of the previous subset. The selection continues until the value of accuracy no longer increases.That also work for backward but the evaluation is when a feature is rejected or when the subset changes. A variation of the criterion is to add a margin of error for the value of the previous subset. This allows to be more flexible and avoid to stop for too small variation in the score. For example, with a margin of one percentile if the value decreases less than 1 percentile compared to the previous subset the selection continues.

Chapter 9

Contribution

It is exist in feature selection(FS) few stopping criteria as wrapper stopping criterion or permutation test combined with mutual information. But some others ways can be explored using statistical or mathematical formula. Some propositions are made to find new stopping criteria that can be used as a threshold in FS. These propositions are based on Hoeffding inequality and permutation test. Classic stopping criteria like the wrapper stopping criterion depend of model performance. If the model gives erroneous information the feature selection is directly impact. Using Hoeffding bound and the permutation test allows to not only depend of the model performance to decide when to stop the selection. When a model is trained without enough samples that impacts the training of the model and it accuracy. Hoeffding bound takes into account the number of features to take a decision and prevents to take decision without sufficiently samples. The permutation test ensures that the score obtained by the evaluation is legit. The score is compared with others scores obtained in the same way but with subsets that contain random values. If more than a specific % of scores obtained with random values gives better result than the not permuted subset this one can be questioned.

9.1 Variation of Hoeffding bound

In the case of feature selection and stopping criteria Hoeffding bound could be used as a threshold to determine if there are enough examples to choose the next feature. To know the number of features necessary to split the node with a probability of $1 - \delta$ the hoeffding bound is used. For FS with wrapper or filter the bound is [0,1] so the range R is 1, n is the number of observation. δG the heuristic that needs to be maximized, is the mean of accuracy or CV scores. f_a is the feature that combined with the previous subset gives the highest scores and f_b be the

second-best feature after seeing n examples. $\overline{G} = \Delta\overline{G}(f_a)$ - $\Delta\overline{G}(f_b) \geq 0$ is the difference between their scores values. Hoeffding bound ensure with probability $1 - \delta$ that f_a is the correct choice if $\overline{G} > \epsilon$.

Another proposition is to base the heuristic on subset of selected features themself and their scores. Instead compare the score of the first (f_a) and second (f_b) best features, compare the current best subset (sub_i) of features with the subset selected at the previous iteration (sub_{i-1}). The Hoeffding bound ensures there are enough samples to ensure that the current subset is better than the previous one. If the evaluated subset is better it is kept and FS continues otherwise the subset is rejected and the FS stops. All parameter for ϵ are the same than for classic Hoeffding.

The last compares the current best subset with a permutation of it. Instead of compare the score of the first (f_a) and second (f_b) best features, compare (f_a) and a permutation of it. The idea is to see if this stopping criterion can make the same job than a stopping criterion based on permutation test but in a faster way. Indeed permutation test need to train the model 1 + m permutations in order to compute the score that determines if continue FS is worth or not. Permutation test takes lots of time and resources to be compute which is not always acceptable.

9.2 Permutation test with induction algorithm

The idea is to use the heuristic that choose feature as the estimated statistic. Indeed the permutation test can be computed with the training, cross validation and delta test scores. For each size of subset the feature selection is done in a classic way and the result is the statistic reference $\hat{\mathrm{O}}$. After the function made the permutation in this way : the column of the selected features is taken and permuted, then the heuristic is reevaluated and stored. This operation show by Algorithm 3 is made m times m being set at 50 for the experimentation. Afterward the percentage of permutation values upper than the reference value is computed. If more than 5 % of values are upper than the reference value, the selection is stop otherwise that continues.

Algorithm 3 Permutation code

```
Set α
Set Number of permutation m
for Each number each size of subset do
  FS process
  compute Ô
  Make m permutation O_k
  PValue = the percentile of value upper than Ô
  if PValue > 5 then
    stop the FS
  end if
end for
```

Chapter 10

Stopping criteria with wrapper experimentation

10.1 Recall,goal and choice of experimentation

In feature selection the goal is to choose a limited number of relevance features. This number is determined by stopping criteria that test subsets of selected features when new feature is added or removed. Several stopping criteria have been chosen for the experimentation to compare results. Stability and reliability of each subset of selected features chosen by the different stopping criteria is evaluated to be compared.

The experimentation compares 6 stopping criteria. There is the classic stopping criteria (CW) of a wrapper that continues to accept new feature as long as adding feature improve the accuracy of the model. A variation of it (WM) that take into account a marginal error variation of 1 percentile. The three variations of stopping criteria based on Hoeffding bound. The first based on the best and second best features (H), the second on subset evaluated and the previous selected subset (H2S) and the last on a permutation of the subset (HP). The 6th is the criterion based on permutation test (P).

10.2 Description of the experimentation

The evaluation of each stopping criteria is made after the selection of the subset to evaluate. The subset is taken and evaluated by function that tests the stopping criteria. For example, if Hoeffding bound considers

there are enough samples to make the right choice, the selection for this criteria continues. When a criteria is not satisfied, the previous subset is saved and the selection for this criteria stop. The p-value of each stopping criteria is computed at the same time for each size of subset excepted for wrapper criteria (CW and WM). After all repetition, the means of the number of features selected for each stopping criteria at each repetition is computed as the variance. The means of the number of features is represented by line on graphics.

Algorithm 4 Wrapper code

```
for Each size of subset do
    FS process
    Take the best subset
    Evaluation of classic wrapper criterion
    if score(sub_i) ≤ score(sub_i − 1) then
        stop for this criteria
    end if
    Evaluation of classic wrapper criterion with margin of error
    if score(sub_i) ≤ score(sub_i − 1) +- 1% then
        stop for this criteria
    end if
    Evaluation of three variation of Hoeffding bound
    Evaluation of Permutation test
end for
```

10.3 Results

With the criterion based on CW the selection is made directly from the value of training accuracy or the cross validation score. First all subsets selected by this criterion take generally more features than with the others shows on figures 10.14. Then subsets are larger if the selection is based on the training accuracy than with cross validation. However with the cross validation, the subset size generally stops at the top of the curve and seems to be closed of the best accuracy on figures 10.1. In regards to the stability on tables 10.16 10.17 10.18, it tend to be very low in the case of CV but high with training. The variation with margin error of 1 percentile seems to take the same or a few more features than without it.

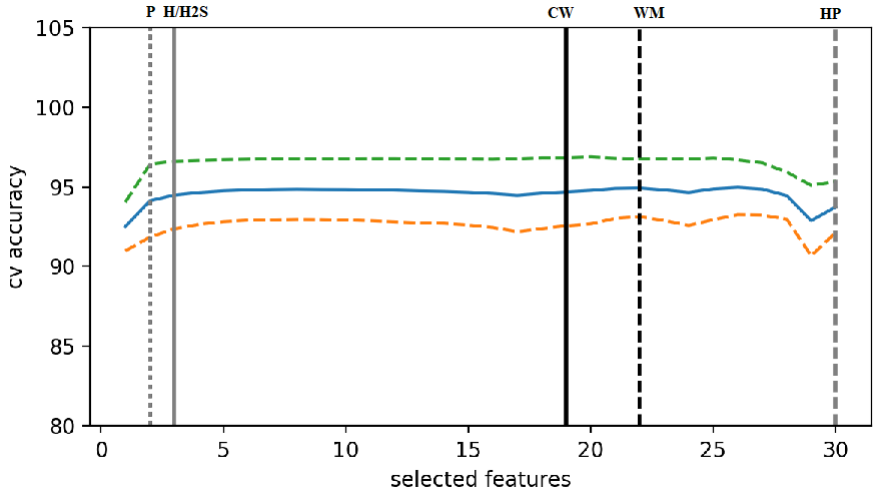
Then with results obtained with the classic Hoeffding the number of features selected is small. whether training accuracy or cross validation

are used the value of the criteria is often very closed in both method. In this case the stability stay high for stability metrics and low on instability metric on figure 10.1 and consequently the stability score is better.

To continue the second version Hoeffding suggests to take more features that the original version that gives a better accuracy for the subset selected even if it is not maximized. Then in the stability graph (10.9), the different kind of stability are high or on little decline. On the color matrix (10.13) the number of features selected often includes the features who stand out.
The last Hoeffding always take all or nearly all features whether it is with training accuracy or CV

Finally the last criteria based on permutation test seem to be the most unstable. The training accuracy gets results closed to the first criterion with subset with too much features. The cross validation show to be closed to the Hoeffding criteria. Furthermore sometimes it is the opposite, in training it will take few features and lot's of features in cross validation.

Breast cancer



Ionosphere

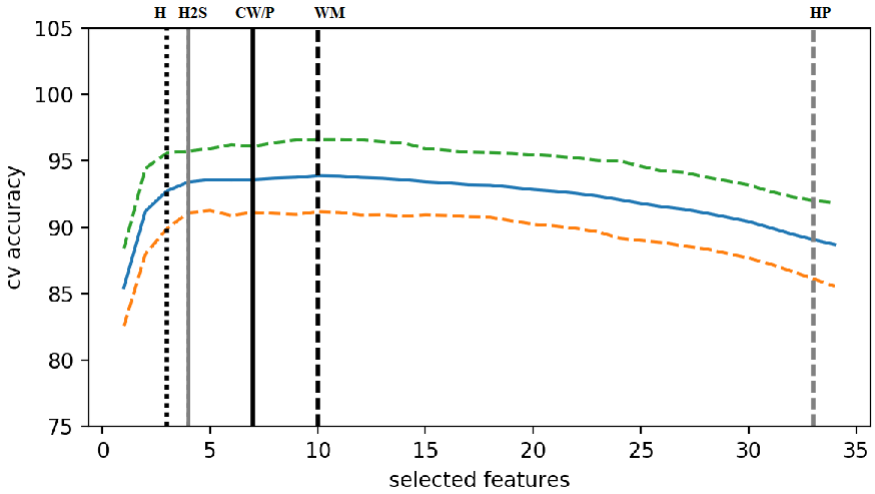
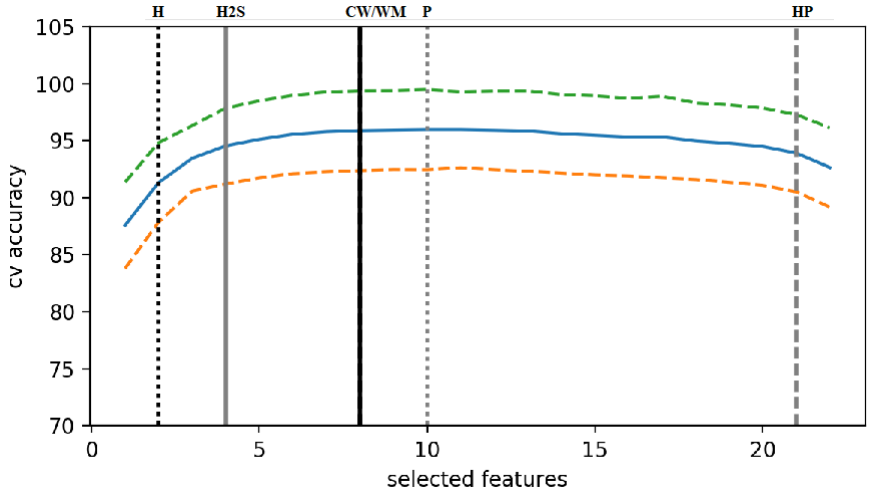


Figure 10.1: CV scores obtained with forward search where selected features is the number of features in the subset. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Parkinson



Sonar

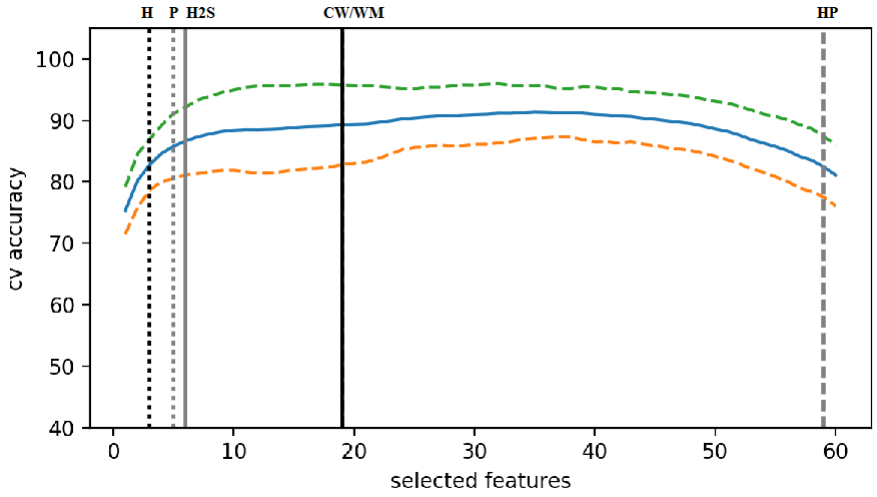


Figure 10.2: CV scores obtained with forward search where selected features is the number of features in the subset. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

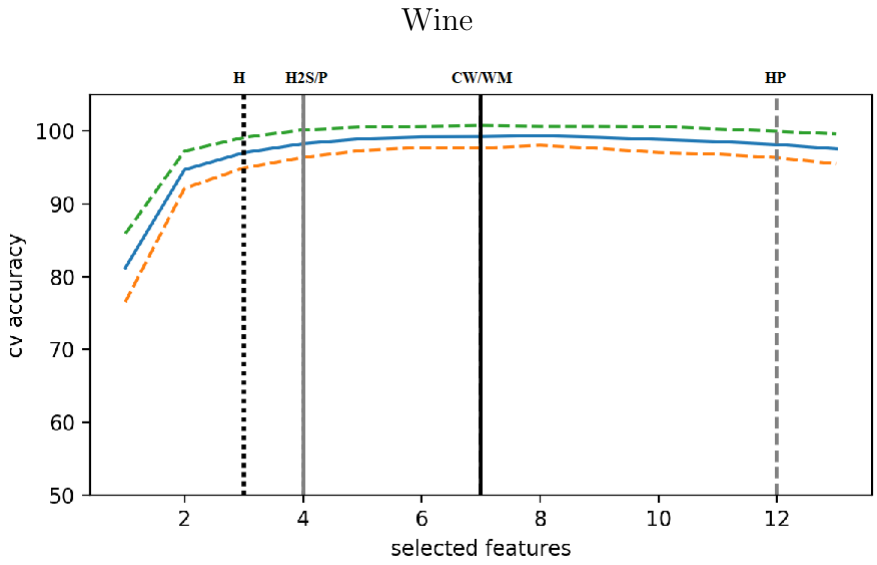


Figure 10.3: CV scores obtained with forward search where selected features is the number of features in the subset. Graphic on the left is experimentation with training accuracy and graphic on the right is experimentation with cross validation score. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

10.4 Discussion

In order to obtain the best features selection it is important to have a subset of selected featured that maximized or approached the maximum of an observed heuristic. Obtain a subset with features selected in a regular and stable way is also important because the goals of FS is to choose specific features and eliminate others. This is done in the optics of reduce the features set which reduces the complexity and the computing time necessary without sacrificing the prediction power of the model.

Base the stopping criterion on the classic wrapper criterion with a FS based on CV is generally closed to the maximized accuracy scores show by figures 10.1. But despite this, the stability associated is often too low to be trust and random features risks to impact the learning phase of the model. Color matrix 10.13 shows the same observation with the margin between stable and unstable that is exceeded and shows that a part of the selection is taken randomly. Furthermore, the variability in the selection (10.15) is very high for this selection criteria. The margin of error of 1 percentile is not usable because the same number of features or more are generally taken and thus worsened the problem

The permutation test is the most unstable of all stopping criteria. Training accuracy shows on graphs 10.1 are closed to the first criteria and obtains subset with too much features to be used. But the cross validation seems to get something closed to the Hoeffding criteria. Furthermore sometimes it is the opposite, in training it takes few features and lots of features in cross validation. There is not pattern with the stopping criteria based on permutation, it can have better results in some cases and awful one in others situations. In the case of this experimentation this unsuitability causes the criterion to be rejected because the features selection can not rest on random results.

Concerning Hoeffding bound, the classic version of the algorithm have a very good stability in the case of CV but the number of features chosen is very small which affect the prediction power. Graphics 10.1 show that the accuracy links to the number of features chosen with classic Hoeffding is usually under the maximum accuracy that can be obtained. The variation of Hoeffding that compares current and previous subset of selected features gives better results. The accuracy obtained with subset chosen by this variation is often closed to the maximum. Stability scores

Breast cancer

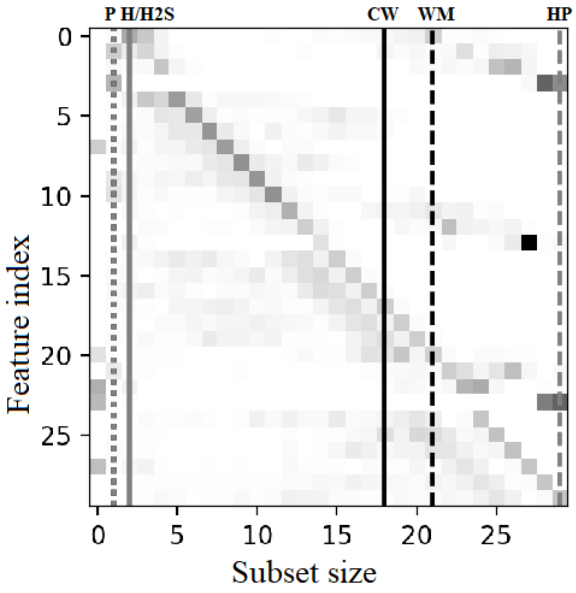


Figure 10.4: Color matrix obtained with forward search base on CV value.

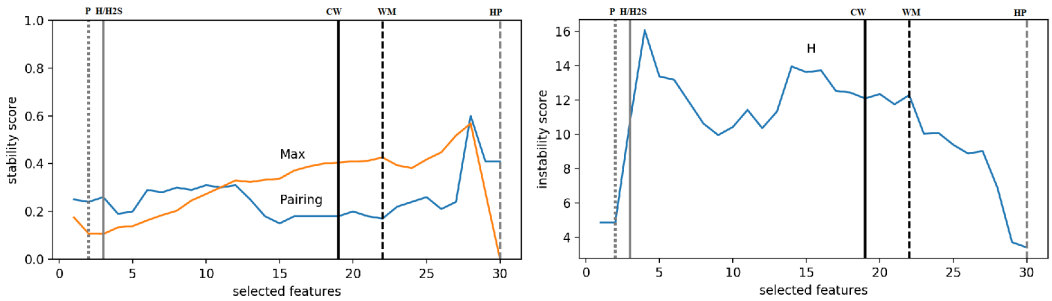


Figure 10.5: Stability graphics obtained with forward search base on CV value where selected features is the number of features in the subset. Graphic on the left is stability score where the higher is the better and the graphics on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Ionospehre

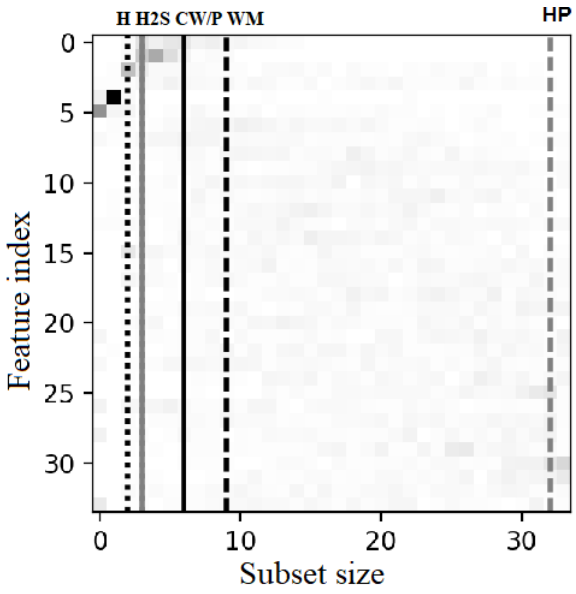


Figure 10.6: Color matrix obtained with forward search base on CV value.

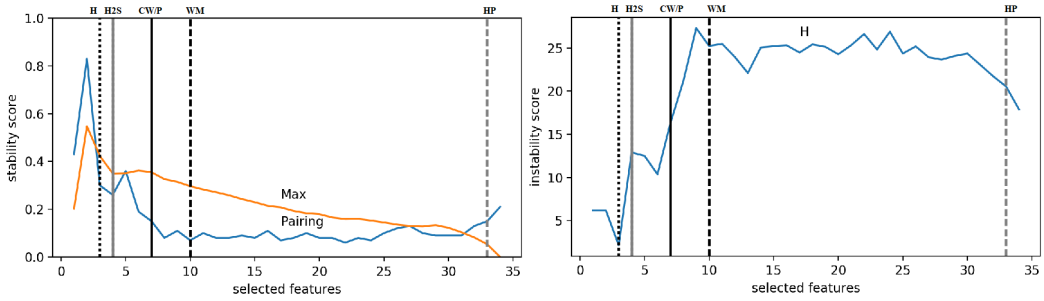


Figure 10.7: Stability graphics obtained with forward search base on CV value where selected features is the number of features in the subset. Graphic on the left is stability score where the higher is the better and the graphics on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Parkinson

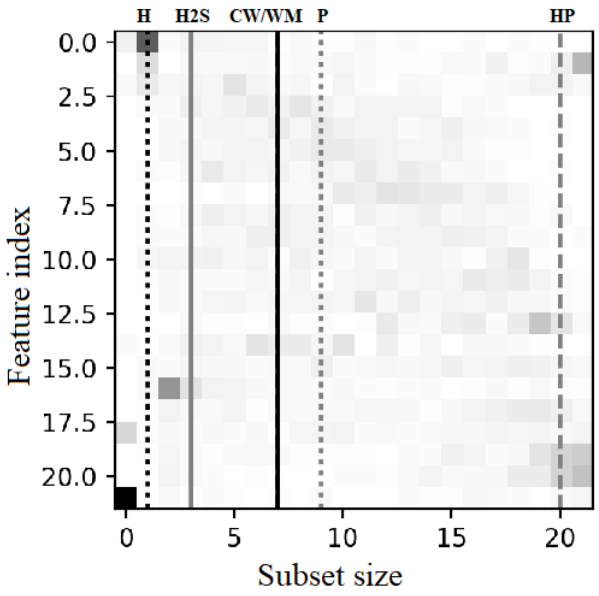


Figure 10.8: Color matrix obtained with forward search base on CV value.

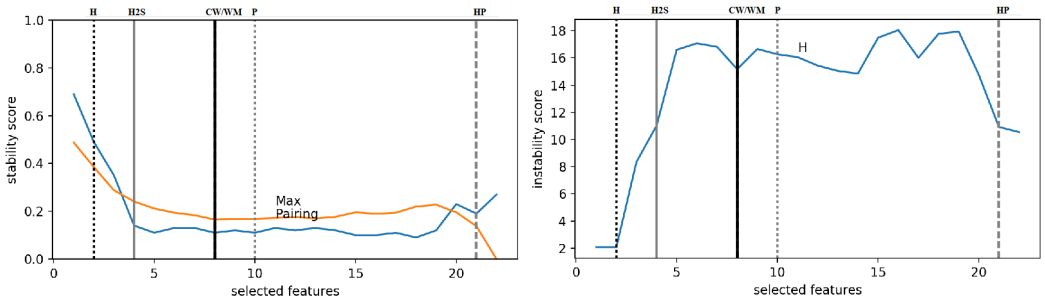


Figure 10.9: Stability graphics obtained with forward search base on CV value where selected features is the number of features in the subset. Graphic on the left is stability score where the higher is the better and the graphics on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Sonar

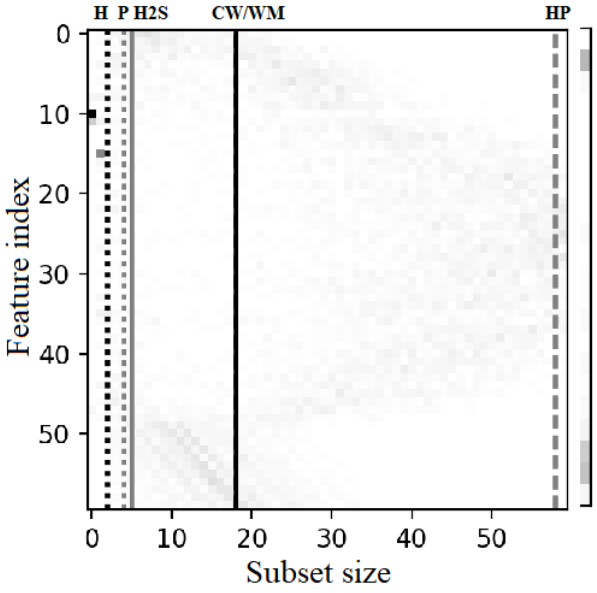


Figure 10.10: Color matrix obtained with forward search base on CV value.

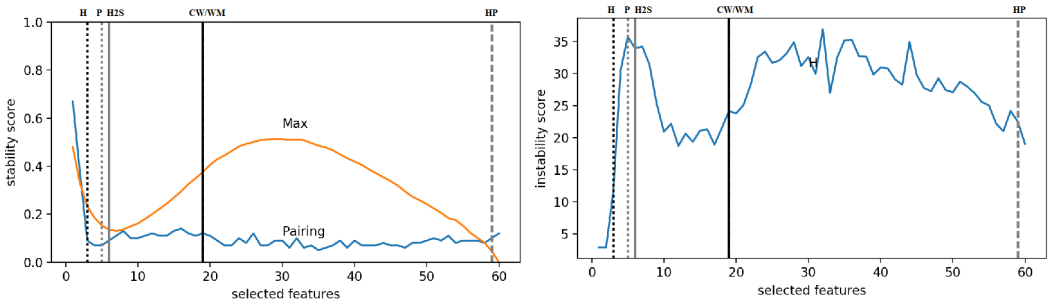


Figure 10.11: Stability graphics obtained with forward search base on CV value where selected features is the number of features in the subset. Graphic on the left is stability score where the higher is the better and the graphics on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

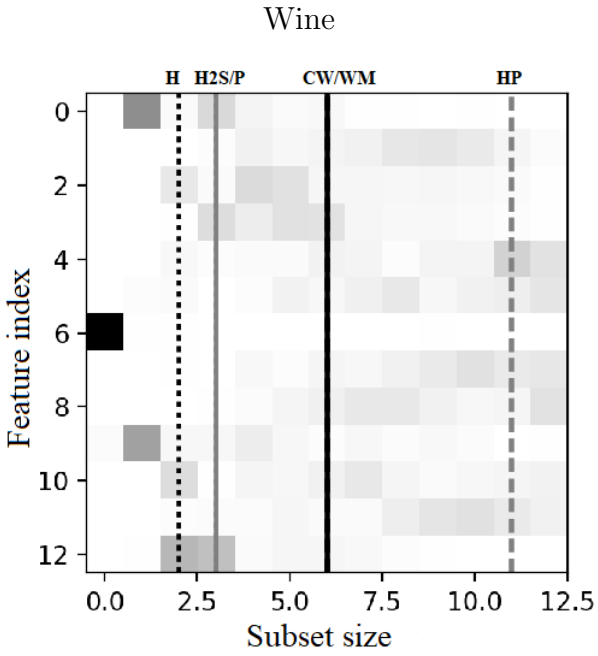


Figure 10.12: Color matrix obtained with forward search base on CV value.

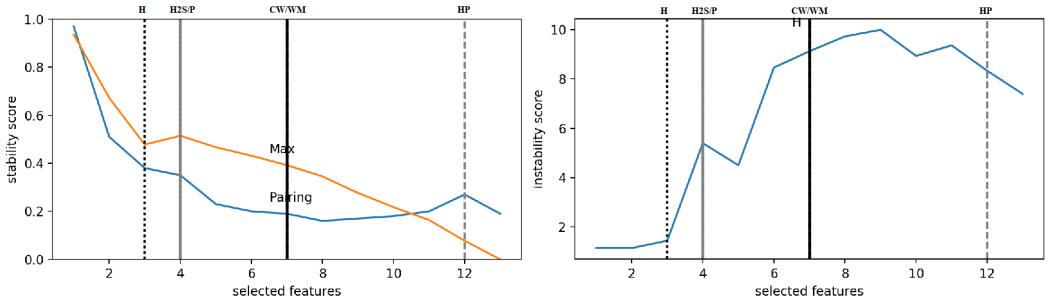


Figure 10.13: Stability graphics obtained with forward search base on CV value where selected features is the number of features in the subset.. Graphic on the left is stability score where the higher is the better and the graphics on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Subject	PM	PMM	H	HS	HP	P
Breast	19	22	2	3	30	2
Iono	7	10	3	4	33	4
Parkinston	8	8	2	4	21	10
Sonar	19	19	3	6	59	5
Wine	7	7	3	4	12	4

Figure 10.14: Number of feature selected with CV

Subject	PM	PMM	H	HS	HP	P
Breast	92.75	60.09	1.37	1.69	0.22	0.95
Iono	6.3664	39.49	5.35	2.04	0.13	1.67
Parkinston	11.93	11.93	2.11	1.85	0.0384	26.70
Sonar	42.08	42.08	3.45	3.90	0.06	3.87
Wine	4.86	4.86	2.5304	1.12	0.03	1.45

Figure 10.15: Variance (standard deviation) in the number of feature selected by the FS with CV

of the selection are very interesting and on the color matrix 10.13 features choice are often in the stable part of the matrix even if sometime it exceeds. The last variation based on permutation gives unusable results and always take the majority of features which is perfectly useless in the context of feature selection. The under goal that is to find a faster alternative of the permutation does not give expected results.

Subject	PM	PMM	H	HS	HP	P
Breast	12.08	12.28	4.85	10.73	3.40	4.85
Iono	16.29	25.21	2.16	12.90	20.52	12.90
Wine	9.12	9.12	1.44	5.39	8.33	5.39
Sonar	24.16	24.16	11.66	33.95	22.49	35.79

Figure 10.16: Stability entropy

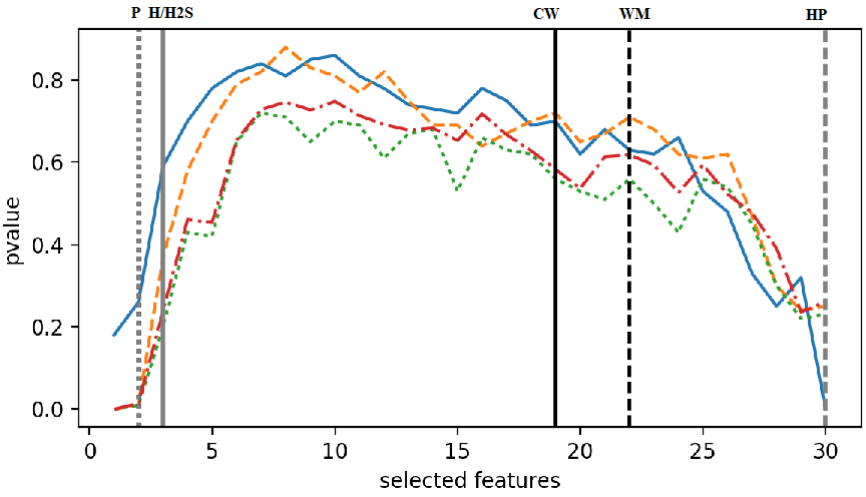
subject	PM	PMM	H	HS	HP	P
Breast	0.40	0.42	0.10	0.10	0.0	0.10
Iono	0.35	0.29	0.42	0.34	0.05	0.34
Wine	0.39	0.39	0.47	0.51	0.07	0.51
Parkinston	0.16	0.16	0.38	0.24	0.13	0.16
Sonar	0.37	0.37	0.23	0.13	0.04	0.15

Figure 10.17: Stability distance

Subject	PM	PMM	H	HS	HP	P
Breast	0.18	0.17	0.24	0.26	0.41	0.24
Iono	0.15	0.07	0.3	0.26	0.15	0.26
Wine	0.19	0.19	0.38	0.35	0.27	0.35
Parkinston	0.11	0.11	0.49	0.14	0.19	0.11
Sonar	0.12	0.12	0.09	0.09	0.1	0.07

Figure 10.18: Stability maximum

Breast cancer



Ionosphere

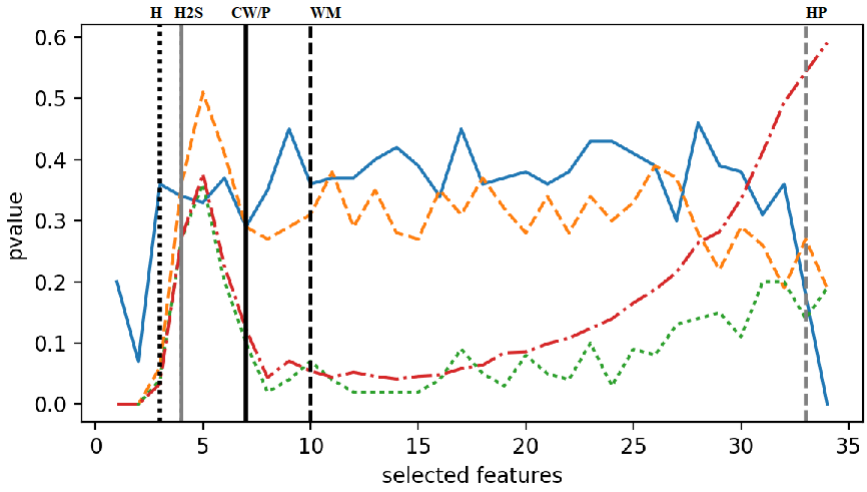
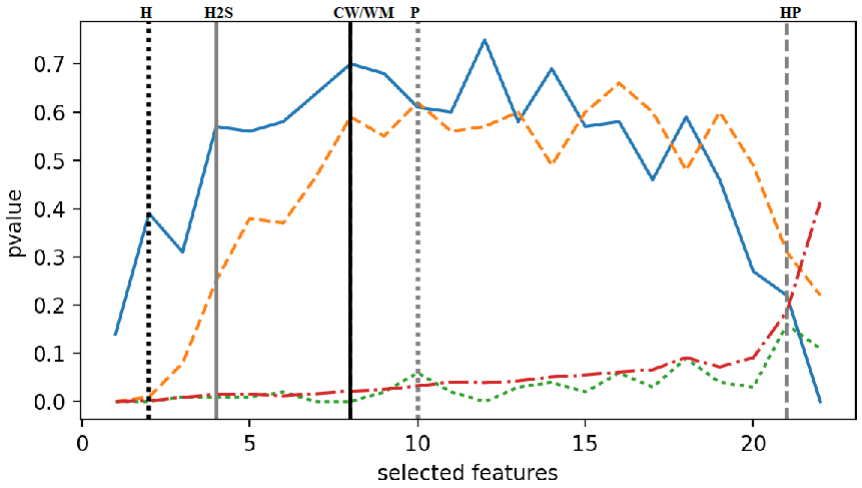


Figure 10.19: *p*value obtained with forward search base on CV value where selected features is the number of features in the subset. Where -(blue) is the Pvalue of classic hoeffding, –(orange) is the Pvalue of hoeffding S, :(green) is the pvalue of hoeffding P and .-(red) is the value of permutation test

Parkinson



Sonar

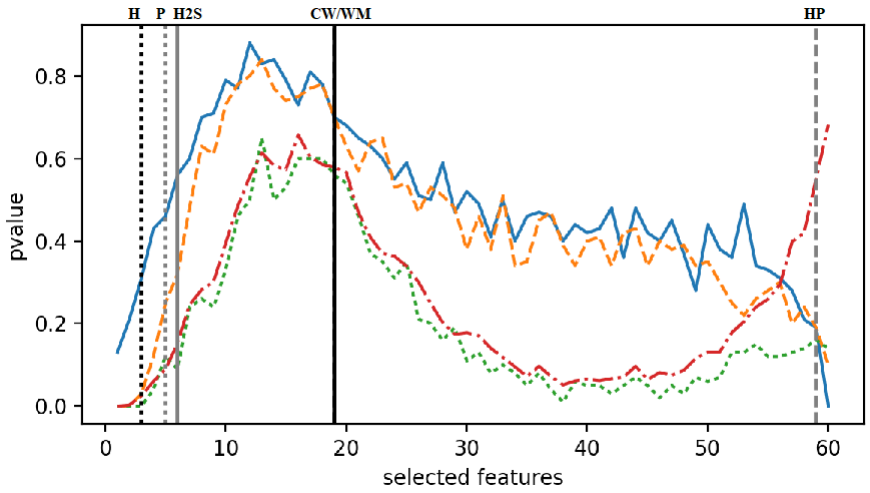


Figure 10.20: *p*value obtained with forward search base on CV value where selected features is the number of features in the subset. Where -(blue) is the Pvalue of classic hoeffding, –(orange) is the Pvalue of hoeffding S, :(green) is the pvalue of hoeffding P and .-(red) is the value of permutation test

Wine

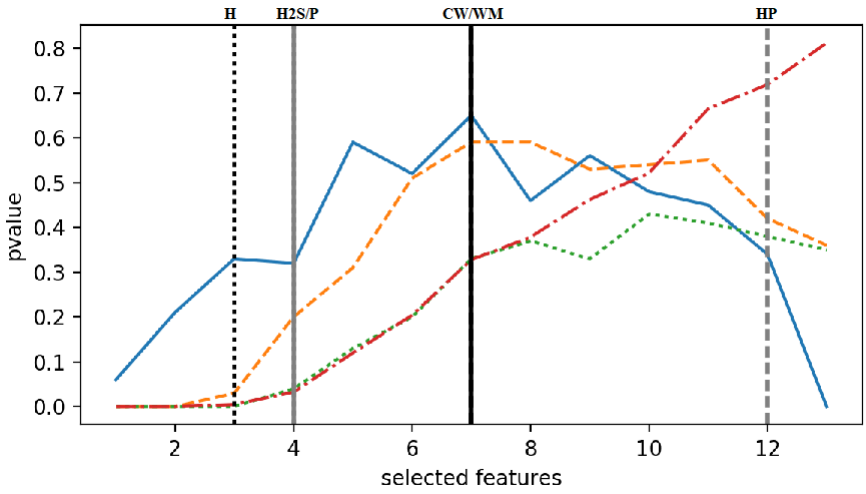


Figure 10.21: *p*value obtained with forward search base on CV value where selected features is the number of features in the subset. Where -(blue) is the Pvalue of classic hoeffding, –(orange) is the Pvalue of hoeffding S, :(green) is the pvalue of hoeffding P and .-(red) is the value of permutation test

Chapter 11

Stopping criteria with filter Experimentation

11.1 Recall,goal and choice of experimentation

As for wrapper the experimentation compares 6 stopping criteria. There is the classic stopping criteria (CW) of a wrapper that continues to accept new feature as long as adding feature improve the accuracy of the model. A variation of it (WM) that take into account a marginal error variation of 1 percentile. The three variations of stopping criteria based on Hoeffding bound. The first based on the best and second best features (H), the second on subset evaluated and the previous selected subset (H2S) and the last on a permutation of the subset (HP). The 6th is the criterion based on permutation test (P).

11.2 Description of the experimentation

The evaluation of each stopping criteria is made after the selection of the subset to evaluate. The subset is taken and evaluated by function that tests the stopping criteria. The p-value of each stopping criteria is computed at the same time for each size of subset excepted for wrapper criteria (CW and WM). After all repetition, the means of the number of features selected for each stopping criteria at each repetition is computed as the variance. The means of the number of features is represented by line on graphics.

Algorithm 5 Filter code

```
for Each size of subset do
    Filter method to select the best subset
    FS is performed
    Evaluation of classic wrapper criterion
    if score(sub_i) ≤ score(sub_i − 1) then
        stop for this criteria
    end if
    Evaluation of classic wrapper criterion with margin of error
    if score(sub_i) ≤ score(sub_i − 1) +- 1% then
        stop for this criteria
    end if
    Evaluation of three variation of Hoeffding bound
    Evaluation of Permutation test
end for
```

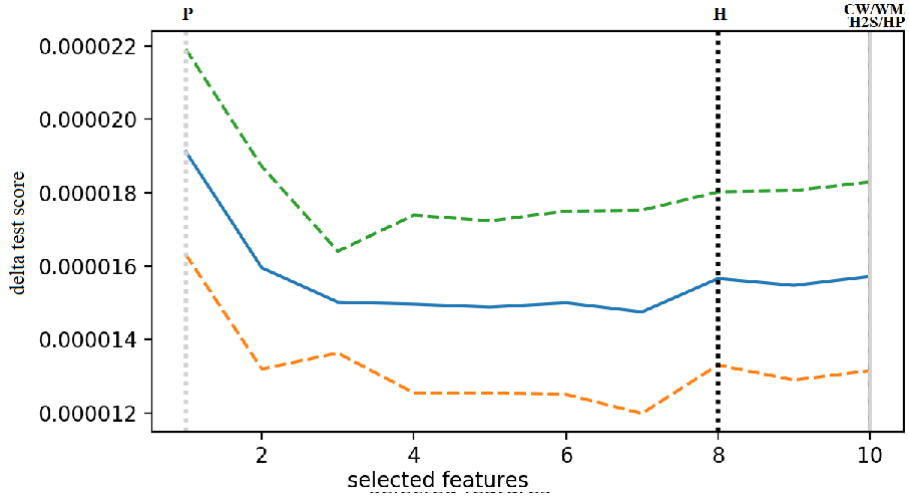
11.3 Results

Stopping criteria have been report on graphics 11.2 11.3 11.4. Criteria based on classic wrapper, classic wrapper with error margin, hoeffding based on subsets and permutation have same results and selects all features. Value of accuracy and stability are low for all these criteria.
The classic Hoeffding criterion takes less features than metrics cited before. The test accuracy for the criterion is good in all case but the stability score is low.
The criterion that uses permutation test take only one feature in all case. The score of stability is high but the score of accuracy is low.

11.4 Discussion

One purposed of the experimentation is to see if observations made for wrapper are also observable for the filter method. With filter stopping criteria seems not work as well as with wrapper. Criteria based on classic wrapper, classic wrapper with error margin, Hoeffding based on subsets and permutation have same results and selects all features. Takes all features goes against the principle of FS because the subset of feature is not reduce. These criteria do not work with filter method using delta test.
The classic Hoeffding criterion takes less features than metrics cited before. The test accuracy for the criterion is good in all case but the stability score is low. The selection select randomly features which brings

Diabetes



Boston

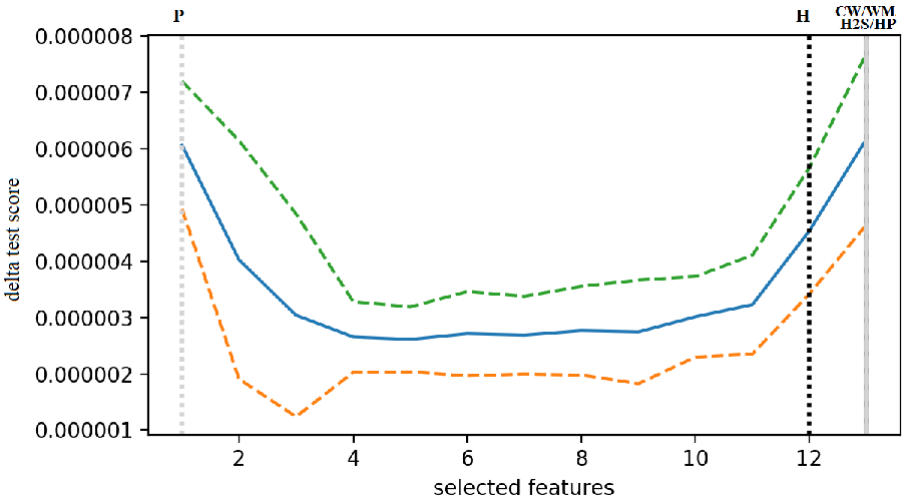
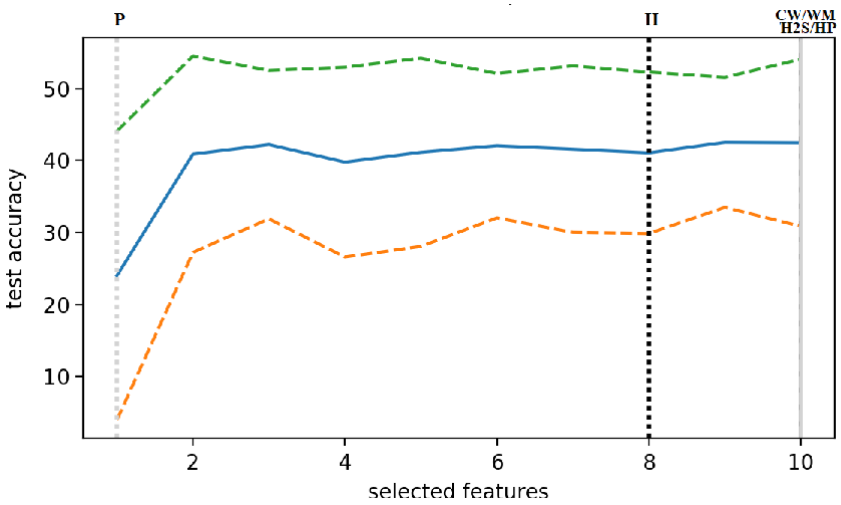


Figure 11.1: Delta test score obtained with filter method where selected features is the number of features in the subset. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P).

Diabetes



Boston

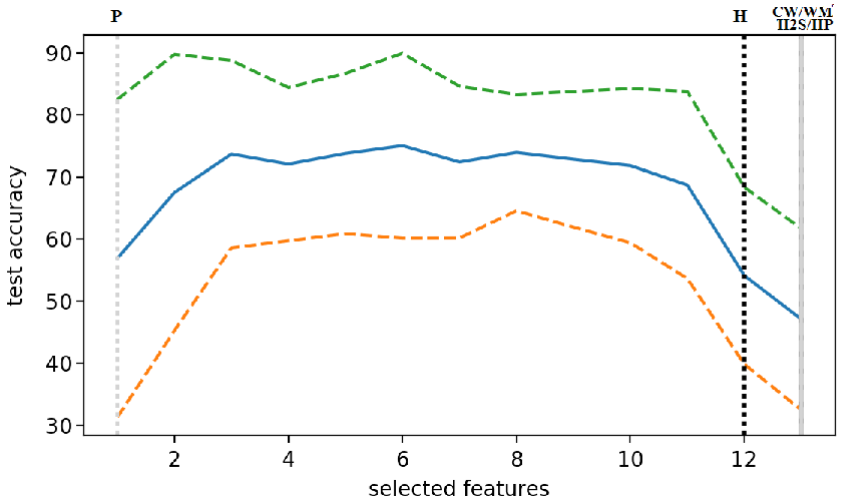


Figure 11.2: Test accuracy score (with R2) obtained with filter method where selected features is the number of features in the subset. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P).

Diabetes

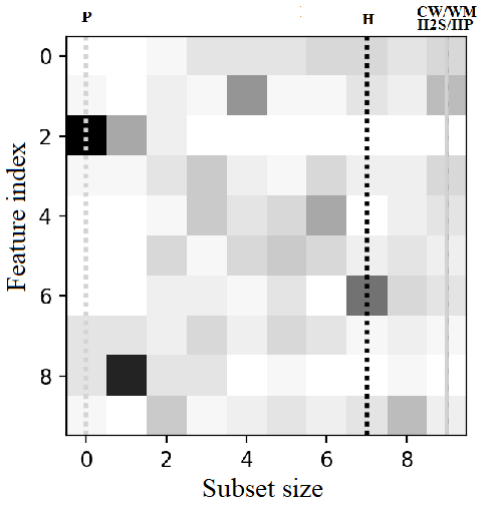


Figure 11.3: Color matrix obtained with filter method based en cv delta score. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

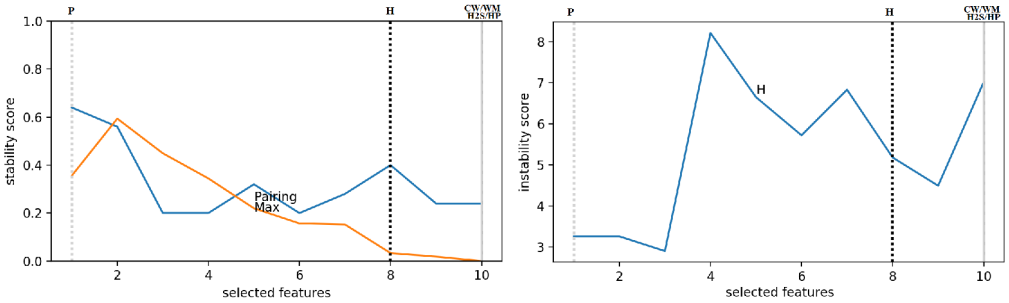


Figure 11.4: Stability graphics obtained with filter method based en cv delta score where selected features is the number of features in the subset. Graphic on the left is stability score where the higher is the better and the graphic on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Boston

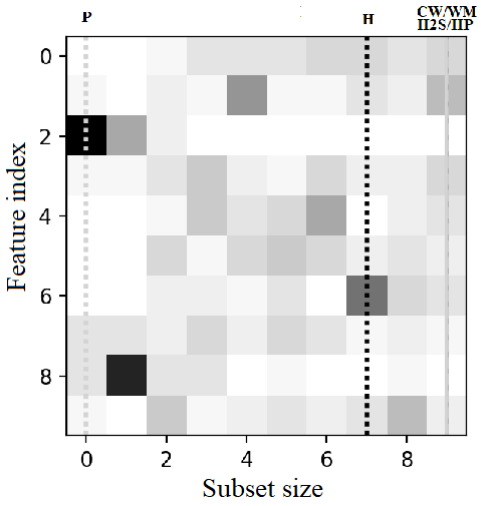


Figure 11.5: Color matrix obtained with filter method based en cv delta score. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

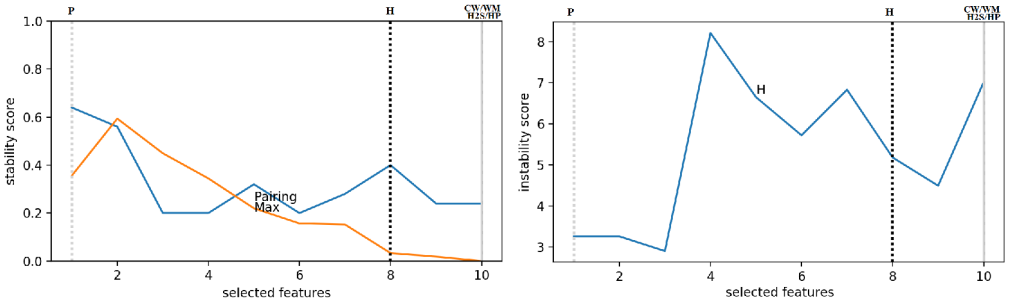
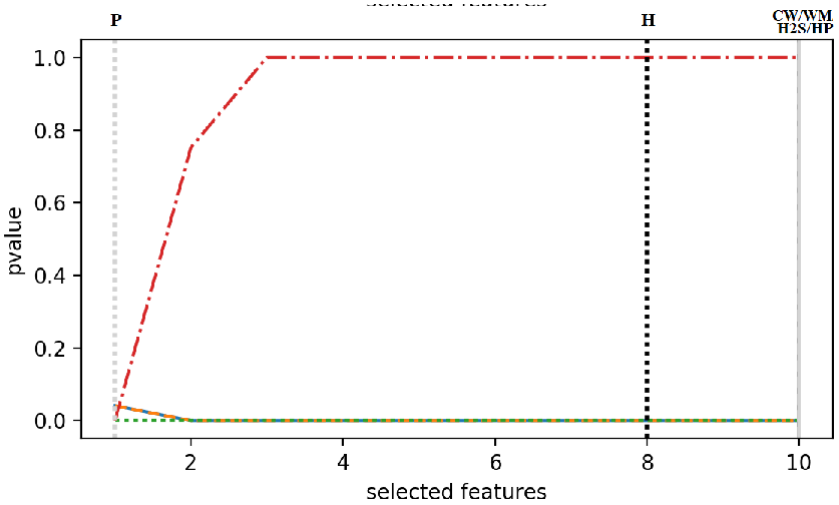


Figure 11.6: Stability graphics obtained with filter method based en cv delta score where selected features is the number of features in the subset.. Graphic on the left is stability score where the higher is the better and the graphic on the right is instability where the lower is ther better. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

Diabetes



Boston

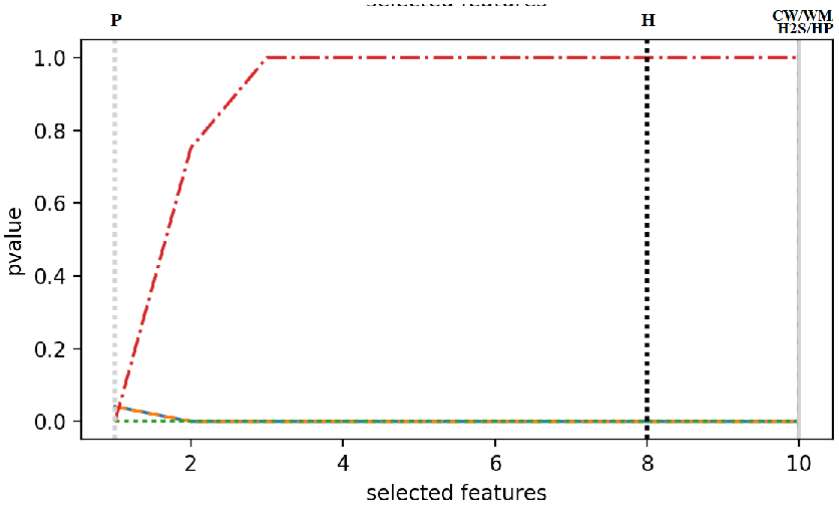


Figure 11.7: Pvalue scores obtained with filter method based en cv delta score where selected features is the number of features in the subset. Where Black - is classic wrapper stopping criterion (CW), black – is classic wrapper stopping criterion with margin (WM), black .. is Hoeffding classic (H), grey - is second Hoeffding proposition (H2S), grey – is Hoeffding with permutation (HP) and grey .. is permutation (P)

nothing to train a model. FS is used to highlight specific features to train the model.
The criterion that uses permutation test take only one feature in all case. The score of stability is high but the score of accuracy is low. Have a low accuracy impact the prediction power of the model. Which is less powerful and useful to perform machine learning.
Regardless of the criterion results show that they are not relevant to perform FS in the case of filter method with delta test.

Chapter 12

Conclusion

12.1 Recall of the goal of the experimentation

The experimentation, on the hand compares feature selection based on training accuracy and cross validation. On the other hand compares proposed stopping criteria adapted from others mathematical forms. The purpose is to test and compare results of each aspect of feature selection made by wrapper and filter method. The comparison is made from results obtained on various datasets and with different kinds of metrics.

12.2 Recall of the principal results

The first research question is focused on selection with training accuracy and cross validate accuracy in the case of a wrapper. The result of the experimentation is that the training accuracy is discouraged to select features. The cap of 100 percentile is too fast reached due to over fitting and that negatively impacts the selection. In fact all features are perceived as worth for the wrapper that makes a default selection by taking each time the first feature still unused . Base the selection on CV gives better results and shows the existence of a border between features that are usually selected and others that are randomly selected.

The second research question is focused on proposed stopping criteria. The second variation of Hoeffding based on the previous subset, has the best result in combination with the selection based on cross validation. The accuracy score for the chosen size of subset is the most of time closed to the maximum. Furthermore the subset seems to take features

highlight by the color matrix. The classic Hoeffding and classic wrapper stopping criterion gives interesting results but only for one metric. That is a problem for the feature selection because one gives a very stable selection but a poor accuracy score and the other gives the opposite. The others stopping criteria give irrelevant or too unstable results to be taken into account in the context of features selection.

The last research question is to see if observation made for wrapper are the same for filter method. Observations are the same for the comparative between selection based on training scores and CV scores. For filter method using delta test, the selection of feature with a CV delta test gives better results than with a delta test computed only with training data. The difference is less visible but the test accuracy and delta test score are better with CV.
About stopping criteria, these do not work well with filter method. They tend to take too much features and not to be balanced in terms of stability and accuracy.

12.3 Future work

Several experimentations can be done to continue the research on feature selection. These experimentations can follow the current goal but also use these results to experiment others ways.

In this experimentation the selection was based on several criteria. Accuracy scores and cross validation scores for the wrapper and $R2$ score for the filter method. But there are others criteria that can be used for the comparison as mutual information. Make experimentation to compare the cross validation that give better results with others selection criteria could be interesting. For the filters there already exist some study focus on the comparison between them.

In the experimentation only wrapper method and filter method as been tested and compared. But there is also the embedded method that can be experimented and compared with the others ones. While still keeping the same purpose than experiment here, embedded method can propose new interesting results. Embedded method use both filter and wrapper method. That can be interesting to optimize the process of selection by using both method. Filter method to reduce the number of features before training the model and wrapper method for robust evaluation could be a good deal. Comparing results of different king of methods to obtain a more accurate result can be another possibility of

experimentation.

Another experimentation which deviates a little from the goal of this paper would be to get interested in features selected. For both wrapper and filter with cross validation, some features are many time selected. Take an interest in these features can be interesting to support experimental results. The number of feature chooses with the second Hoeffding criterion is it optimal ? Why some features are always selected and does that make sense ? There are plenty questions that can be experimented. Focus on features selected them self could be interesting to know why this features are always selected and is it possible to plan which features is choose and which feature to remove.

Bibliography

[1] Ron Kohavi and George H. John b
Wrappers for feature subset selection. Journal : Elsevier: Artificial intelligence
Year : 1997
Pages 273-324
Volume 97

[2] Sebastián Maldonado, Richard Weber * *A wrapper method for feature selection using Support Vector Machines.* Journal : Elsevier: Informations Sciences
Year: 2009
Pages 2208-2217
Volume 179

[3] Noelia Sanchez-Marono, Amparo Alonso-Betanzos and Mara Tombilla-Sanroman *Filter methods for feature selection. A comparative study* Year : 2007
Intelligent Data Engineering and Automated Learning - IDEAL 2007
United Kingdom
Month : December

[4] Vijayawada, Andhra Pradesh *Feature Selection using ReliefF Algorithm* Journal : International Journal of Advanced Research in Computer and Communication Engineering
Year : 20014
Page ?
Volume 3

[5] Marko Robnik, Igor Kononenko *Theoretical and Empirical Analysis of ReliefF and RReliefF* Journal : Machine learning
Year : 2003
Pages 23-69
Volume 53

[6] Sanmay Das *Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection* Year : 2001
ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning
USA
Pages 74-81
Month : July

[7] Pedro Domingos and Geoff Hulten *Mining HighSpeed Data Streams.* Year : 2000
Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining
Boston
Pages 71-80
Month : August

[8] D.Francois, V.Wertz and M.Verleysen *The permutation test for feature selection by mutual information* Year : 2006
Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)
Bruge, Belgium
Pages 239-244
Month : April

[9] Bo Xin∗, Lingjing Hu†, YizhouWang∗ andWen Gao∗ *Stable Feature Selection from Brain sMRI.* Year : 2015
Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence
Bruge, Belgium
Pages ? Month : January

[10] David Dernoncourt a,b,∗, Blaise Hanczar c, Jean-Daniel Zucker a,d *Analysis of feature selection stability on high dimension and small sample data.* Journal : Elsevier
Year : 2014
Pages 681-693
Volume : Computational Statistics and Data Analysis

[11] Ludmila I. Kuncheva *A STABILITY INDEX FOR FEATURE SELECTION.* Year : 2007
International multi-conference ARTIFICIAL INTELIGENCE AND APPLICATIONS
Innsburk, Austria

Pages 390-395
Month : February

[12] Alexandros Kalousis, Julien Prados, Melanie Hilario *Stability of Feature Selection Algorithms a study on high dimensional spaces.* Journal : Knowledge and Information Systems
Year : 2007
Pages 95-116
Volume 12

[13] Sarah Nogueira and Gavin Brown *Measuring the Stability of Feature Selection.* Journal : Machine learning and Knowledge in Databases
Year : 2016
Pages 442-457
Volume 9852

[14] Sarah Nogueira and Gavin Brown *Measuring the Stability of Feature Selection with Applications to Ensemble Methods.* Year : 2015
International Workshop on Multiple Classifier Systems
Pages 135-146
Günzburg, Germany
Month : June

[15] Avrim L.Blum, Pat Langley *Selection of relevant features and Examples in Machine Learning.* Journal : ELSEVIER, Artificial Intelligence
Year : 1997
Pages 245-271
Volume 97

[16] F. GUILLAUME BLANCHET,1 PIERRE LEGENDRE, AND DANIEL BORCARD *FORWARD SELECTION OF EXPLANATORY VARIABLES.* Journal : Ecology
Year : 2008
Pages 2623-2632
Volume 9

[17] Thibault Helleputte, Pierre Dupont *Partially Supervised Feature Selection with Regularized Linear Models.* Year : 2009
Proceedings of the 26th Annual International Conference on Machine Learning
Pages 409-416
Montreal, Quebec, Canada
Month : June

[18] Jasmina NOVAKOVIĆ, Perica STRBAC, Dusan BULATOVIĆ *TOWARD OPTIMAL FEATURE SELECTION USING RANKING METHODS AND CLASSIFICATION ALGORITHMS*. Journal : Yugoslav Journal of Operations Research
Year : 2011
Pages 119-135
Volume 1

[19] Jason Van Hulse, Taghi M. Khoshgoftaar,Amri Napolitano, Randall Wald *Threshold-based feature selection techniques for high-dimensional bioinformatics data*. Journal : Network Modeling Analysis in Health Informatics and Bioinformatics
Year : 2012
Pages 47–61
Volume 1

[20] D. Fran¸cois a, F. Rossi b, V. Wertz a and M. Verleysen c, *Resampling methods for parameter-free and robust feature selection with mutual information*. Journal : Neurocomputing 70
Year : 2007
Pages 1276-1288
Volume 7-9

[21] L. Rutkowski, Fellow, IEEE, L. Pietruczuk, P. Duda and M. Jaworski *Decision trees for mining data streams based on the McDiarmid's bound*. Journal : IEEE Transactions on Knowledge and Data Engineering
Year : 2013
Pages 1272-1279
Volume 25

[22] Zhenyun Deng ,Xiaoshu Zhu, Debo Cheng, Ming Zong ,Shichao Zhang *Efficient kNN classification algorithm for big data*. Journal : Neurocomputing
Year : 2016
Pages 143-148
Volume 195

[23] Kilian Q. Weinberger, John Blitzer and Lawrence K. Saul *Distance Metric Learning for Large Margin Nearest Neighbor Classification* Journal : Journal of Machine Learning Research
Year : 2009
Pages 207-244
Volume 10